汉译经济学文库

微观经济计量学

——方法与应用

［美］ A. 科林·卡梅伦（A. Colin Cameron） 普拉温·K. 特里维迪（Pravin K. Trivedi） 著

王忠玉 译

上海财经大学出版社

译者序

当前，微观经济计量学作为经济计量学前沿领域的一个新分支，在最近30多年间得到了迅速发展。微观经济计量学侧重于对家庭、厂商等个体经济行为进行计量研究，其研究范围非常广泛，涉及的专题包括：劳动力供给、工资决定、教育选择、失业持续期限、移民、职业选择、生育选择、性别歧视、种族歧视等劳动经济学专题；税收政策及福利政策的效应等公共财政专题；商品需求、品牌选择等消费行为专题；住所选择、区位选择、交通工具选择等都市及运输经济学专题；生产形式选择、生产要素需求、生产效率评估等产业经济学专题。微观经济计量学几乎涵盖了所有涉及个体经济方面的专题。

这部由卡梅伦和特里维迪所著的《微观经济计量学——方法与应用》，除详细介绍微观经济计量学中广泛运用的各类模型理论基础之外，还特别强调微观经济计量方法的实证应用，突出了对建立及运用模型的过程中可能产生的各种实际问题的处理。而且，作者对有关最新进展专题或特定模型的估计及检验方法进行了逐一评述。实际上，本书几乎囊括了当今微观经济计量学的各类专题。书中内容专题众多，体现出两位作者极高的学术造诣，他们特别擅长统计学方法，在阐述微观经济计量建模问题时，其经济计量建模思想深邃、建模技术娴熟，使你在研读之后真正体会到，统计学方法或数学工具只是进入前沿领域的一块基石，更重要的是拥有一种经济计量建模的理念及直觉力。

正如经济计量学家、美国南加利福尼亚大学萧政教授所赞誉的："这本书对当前微观经济计量学家所广泛研究的迅速发展的专题给出了优美而深入浅出的处理。以富于创见、直观精湛的方式对复杂的经济计量方法论核心概念加以设计，本书不仅对大学生而言是一部优秀的教科书，对实践者与研究者来说，也是一部非常宝贵的参考书。"

(一) 经济计量学到底有什么用途？

这里不能不提及，经济计量学到底有什么用途？换句话说，经济计量学的作用何在？尽管稍稍拥有经济计量学知识的人士对这个问题都可能略知一二，为了对当代经济学有一个深刻的认识和理解，我们有必要在此更深入地考究一番。

经济计量学的用途主要有四个方面：

第一，经济计量学最明显的用途是用于**检验经济理论**的含义是否正确。经济理论的实质是一系列的假设。因此，要检验经济理论正确与否，就是要检验其假设是正确的还是错误的。

已故经济学家米尔顿·弗里德曼(Milton Friedman，1912～2006年)在他的著作《实证经济学方法论》(1953)中认为："理论的实质是一系列的假设……一般而言，可以发现，真正举足轻重的假说的'假设'都是对现实的一种粗略的、不十分精确的描述。而且，通常理论越是重要，其'假设'就越是不真实。个中原因非常简单……因此，从意义重大的角度来看，假说对于假设就不能是忠实的描述；假说对于解释已有现象的成功，表明具体现实环境因素的影响是有限的，那么假说的假设自然也就不必对现实环境亦步亦趋。"

英国经济计量学家亨德瑞(Hendry)认为："经济计量学的三个信条是：检验、检验、再检验。"

第二，经济计量学用于**测算理论上定义的参数或者不可观测变量的未知值**。在极端情况下可以认为，经济计量学是发现经济现象的助推器，即直接测量由经济理论提出的基本关系，譬如柯布—道格拉斯生成函数的创立及发现。

上述两种作用均将理论置于证据之前。这就存在两种情况：其一，先提出一种经济理论，而后整理支持该理论的证据；其二，经济理论对于所给定义或者测算目标来说极其重要。

第三，经济计量学用于**预测变量值**，预测可直接建立在先验经济理论之上，或者预测是一种非理论的统计演算。进行预测时，要假定拥有进行样本外推关系的平稳性。另外，理论说明经常为偶然规律和真实规律之间的差异提供了强有力的支持。如果缺少这一点，那么经济计量研究者就会同股票市场图形研究者没有什么两样。

第四，经济计量学用于**刻画一种经济关系或现象**。经济计量学所包含的数据可以揭示出特定经济变量之间的关系，从而成为支撑理论的素材。

(二) 从经济计量学方法论视角，提升实证分析建模认知

厦门大学王亚南经济研究院院长、经济计量学家洪永淼教授认为："现代经济计量学实际上是建立在以下两个基本公理之上的：(1) 经济系统可以看作服从一定概率法则的随机过程；(2) 经济现象(主要表现为经济数据)可以看作这个随机数据生成过程(data generating process)的实现。"

尽管现代经济计量领域大量运用高等数理统计方法及理论，但这两者之间的方法论仍然是有差异的。更准确地讲，经济计量学方法论与应用于经济中的统计学方法论是有区别的。那么，经济计量学方法论与应用于经济中的统计学方法论的区别在哪里呢？或者说，两者之间的最大差异是什么呢？

众所周知，高等经济计量学广泛运用高等统计学知识。毋庸置疑，具有高等统计学知识为理解、认识和掌握高等经济计量方法论提供一个较好的基础，但在学习及研究高等经济计量学知识时，仍会遇到诸多困难，其原因何在？

统计学方法论就是要正确地揭示概率并说明它是如何应用于数据中的。这

里,关于概率的经典解释和贝叶斯解释之间的争论是一个核心内容。当然,这同样也是经济计量方法论的中心议题。不过,这里将不涉及该议题。

我们认为,要想解决这个普遍存在的问题,就必须提升对经济计量方法论的理解层次,在经济计量学方法论的认知方面有一定提高及进步。为此,首先要清楚地认识和掌握经济计量学方法论与应用于经济中的统计学方法论的区别。这两种方法论的差异存在于两个方面。

第一,英国哲学家南希·卡特赖特(Nancy Cartwright)认为,与社会学不同,经济计量学揭示了统计学应用,"是运用理论的学科"。

卡特赖特的这个观点引起了许多经济计量学家的共鸣,认为经济理论必须为统计经济解释提供所需要的识别。这样便遭遇到所有的先验方法问题:必须拥有正确的理论来定义自然法则机制或者识别模型,但倘若推断方向仅仅从理论到数据,那么怎样运用经验证据来确定哪一种理论是正确的呢?实际上,对此类问题的讨论将涉及科学哲学的诸多观点与流派。

不过,应该提及的是,经济计量学家哈维默(Haavelmo,1911～1999年)在1943年建立了经济计量学的概率论基础。这为近现代的经济计量学进一步发展打下了坚实基础。

纵观经济计量发展史可以发现,严谨的经济计量理论的发展紧密地依赖于统计学的最新进展,经济计量领域的任何重要发展都源于此。早先文献中的许多困惑,通过利用概率工具得以澄清,从而更准确定义与辨析那些易混淆的知识。与此同时,数理统计学的迅速发展、先进成果转移及推广,极大地激发了经济计量理论研究,这类动因无论是在理论深度还是在应用广度上都得到了不断发展。

第二,经济计量学方法论关注于建立因果关系,而统计学通常满足于建立相关关系。诺贝尔经济学奖得主、经济计量学家赫克曼(Heckman)教授认为:"大部分经济计量理论采用的都是最初源自统计学的研究方法。有一个重要的例外,即识别问题的经济计量分析,还有与之相伴的结构方程分析、因果性分析以及经济政策评价。"

"20世纪经济计量学对知识的重大贡献是对因果参数的定义……为了揭示来自数据的因果参数而需要的分析……政策评价的因果参数作用得以澄清。"

经济计量学作为一种因果科学思想,颇为引人注目。可是,几乎可以肯定,这种思想是一种历史观。最近20多年来,不论是微观经济计量研究还是宏观经济计量研究,都在试图恢复揭示因果关系的经济计量建模方法。

(三)经济实证方法的两大派别

通常,经济模型有两大类:一类是理论模型,另一类是实证模型。理论模型是从经济理论中直接导出,而实证模型则是从理论模型衍生出来,要用实际数据来估计。一般来讲,实证模型是以回归模型形式表示,对模型中所涉及的变量均要给予明确定义,并对解释变量和因变量之间的关系详细说明,此外,也要对模型的主要系数或由这些系数导出的弹性可能数值的大小及符号给予一定的预期。

一般来说,对于建立实证模型时如何利用经济理论的问题,不同研究者有着不尽相同的观点,他们可能会产生一些争论,甚至出现截然不同的观点。目前,就这

个问题而言，存在两种极端方法。一种方法认为，理论包含着唯一、纯粹的真理，因而应成为模型基础。持有该观点的研究者声称，所有的残差都应该得到理论的解释，而不给随机性、不确定性或系统的外生冲击留有一席之地。这种建模方法也称为结构方法，认为数据不可能完全显示自己是怎样产生的。结构方法起源于考尔斯委员会。

持有结构方法的研究者认为，假如说经济研究的目标是数据生成过程(DGP)，则只有在研究者模型的协助下才能了解数据产生结构，尽管研究者模型可能是错误的。从科学研究方法看，结构方法非常接近于物理学研究方法。众所周知，物理学家从事科学研究的方法有：(1) 数学理论(主要是数学模型)；(2) (实验室中的)实验方法；(3) 计算机模拟法。物理学家想要了解物质是怎样运转的，通常先提出模型，然后用实验加以检验。物理学家的模型可能是错误的，即使模型与目前所有的数据符合；但倘若没有模型，物理学家的理论就无从运用，因为一大堆无模型的数据不能被用来预测。

持有结构方法的经济研究者注重模型，强调估计模型的原始参数。所谓原始参数是指那些在偏好和技术方程中的参数。这些参数不会因为政策干涉而变化。相反，应用简化方法估计的参数多数不是原始参数，因而无法用来进行预测，尤其无法预测从来没实施过的政策会有什么影响。

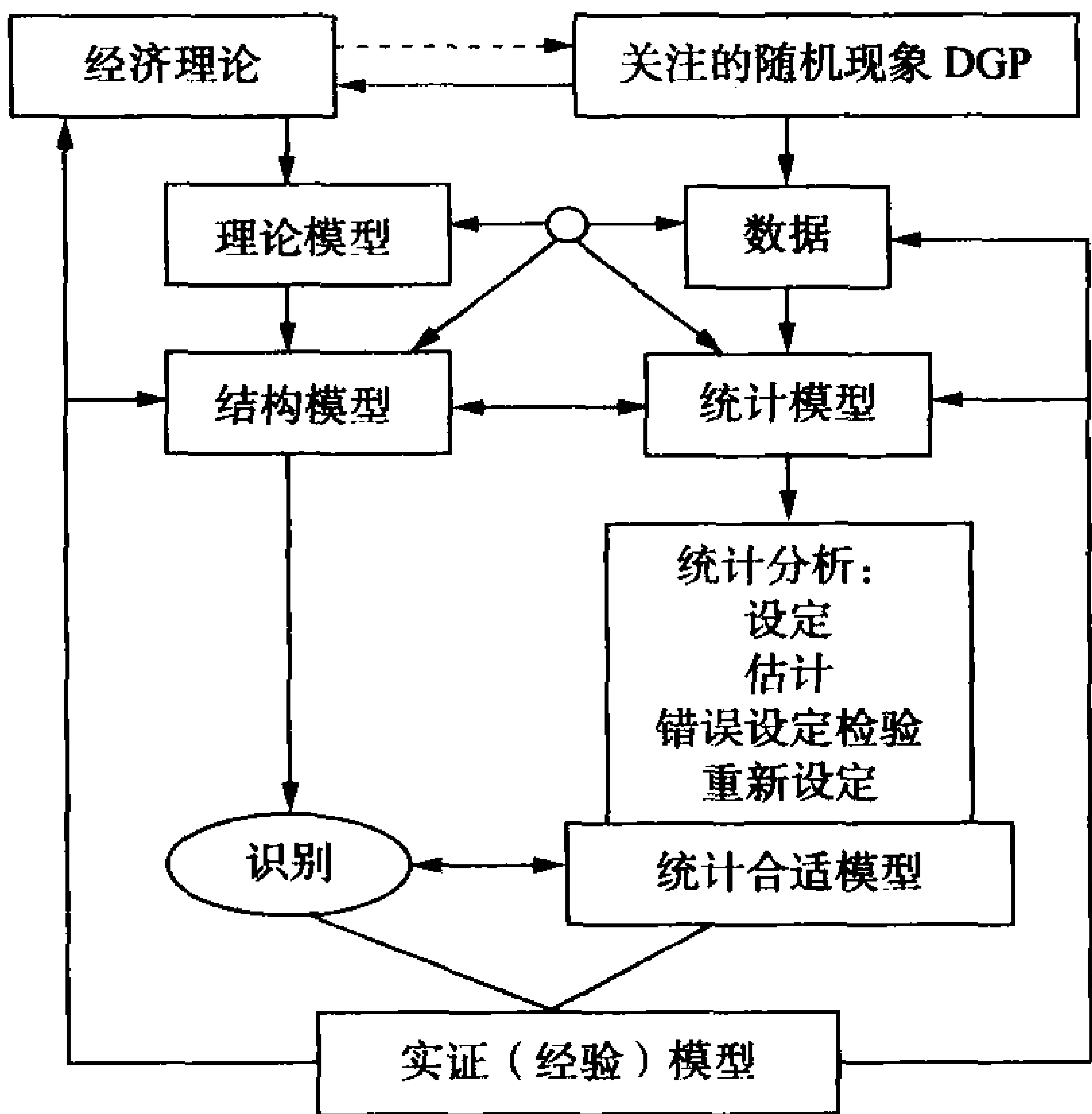
另一种方法认为，只依据经济现象所呈现出的规律性和关联性，建立基于对数据观测的“非理论的”模型。这种建模方法称为简化方法。简化方法与结构方法的区别在于它们对经济理论在实证研究的作用定义不同。

简化方法认为，实证研究应该让“数据自己说话”，认为经济理论模型是由研究者意志决定的，将研究者的认识和看法施加到数据上而得到的结论，只有在模型正确的情况下才会正确。由于研究者不可能知道什么模型是正确的，他们的主要研究工具很简单：使用各种各样的回归分析。实际上，实证研究不仅仅看数据，更重要的一部分内容是对实证研究方法的研究，因为任何一篇以数据为主的论文，其研究结果的有效性取决于所用的研究方法。一个很好的例子是，赫克曼的选择模型完全改变了之前关于劳动力供给的研究成果。

对上述方法论的认识，可用一个“经济理论—经济数据—建模方法—实证模型(或经验模型)”的全程建模框架来刻画及描述，如下页图所示。

总之，不论是学习、研究高等经济计量学理论及方法，还是提升自己对经济问题的计量建模水平和认知层次，除了扩大特定的前沿专业知识外，还应该围绕经济计量建模方法论框架展开。对经济计量建模全过程，要有一个清晰的认识和理解，做到知方法、明过程，促使自己今后在探究经济问题时更有章可循，其终极目的就是用规范的实证研究方法表达新的观点，或者阐明对经济现象内在规律的发现。

译者先后参与 2007 年教育部重大项目(07JJD790131)、2008 年教育部重大项目(08JJD790153)、“吉林大学‘985 工程’项目”经济分析与预测创新基地项目(985CXJD001)的研究工作，以及作为跟踪和掌握经济计量领域最新进展的翻译和研究工作，为理论基础及方法创新起到良好的支撑作用，并在上述各个项目的探索和研究中发挥了重要的学术作用，提供了重要的参考价值。



经济计量建模方法论框架图

译者在翻译本书过程中，曾得到经济计量学家、美国 Portland State 大学林光平教授、吉林大学数量经济研究中心赵振全教授的帮助和支持，还得到了哈尔滨工业大学王雪松副教授、博士研究生孙薇，以及硕士研究生朱晓燕、徐婉和夏义星同学的支持，为此译者衷心地感谢他们。另外，还要感谢哈尔滨工业大学经济与管理学院院长于勃教授对译者从事数量经济学教学及科研工作的关怀、勉励，这里要特别感谢冯英俊教授的多年教导、支持和鼓励，并提议将本书的部分内容用于《经济数学模型与方法》博士生课程加以讲授。此外，也要特别感谢妻子与儿子的大力支持，他们的默默奉献一直激励着我从事艰苦的翻译工作。

最后，感谢上海财经大学出版社的袁敏先生富有创新、超前视野的工作理念与支持，以及编辑温涌女士耐心细致、兢兢业业的工作，她的文字润色使得本书译稿更加流畅。

值得提及的是，原书作者卡梅伦教授对于译者翻译本书给予了诸多指导，使得译者对本书的理解更加深刻、翔实。在对这样一本厚重著作的专业翻译工作中，难免出现纰漏及错误，在此恳请专家和读者指正。译者的联系方式为：wangzhy@hit. edu. cn。原书中的一些印刷错误，译者已经逐一做了改正。

王忠玉

2010 年 6 月 25 日

序　言

本书针对微观经济计量分析做出了详细研究，内容涉及对揭示个体或厂商经济行为的个体层面数据加以分析。此类分析通常需要对横截面与面板数据运用回归分析方法。

本书旨在为应用研究者提供一种综合的统计方法，以及将其用于现代微观经济计量领域的研究方法。这些方法包括：非线性建模方法、最小分布假设条件下的推断、识别与测量因果关系而非纯粹的关联，以及对违背简单随机抽样加以修正。在社会科学中，这些特性全部与个体层面数据分析有关。

如此雄心勃勃的设想决定了本书的特点。第一，本书虽然是面向应用研究者的，但在层次上属于相对高级的水准。由于两种以上的因素同时发生作用是一种常见情况，所以采用照搬手册的方法明显不合适，因此，应用研究者必须掌握足够多的知识，以便采用合适的方法。第二，本书给出相当多的实际数据问题（尤其在最后三章）。第三，为了阐明所述方法，本书许多章节包括大量的实证例子。最后，本书篇幅非常长。尽管在篇章上我们加以限制，但仍旧是非常厚重。我们在分析时包括特别多的实证例子，倘若运用简略描述，经常会使研究者做出的实质贡献无法被揭示出来。

本书假定读者能理解用矩阵代数形式表述的线性回归模型。与格林（Greene，2003）的书相比，本书使用的数学知识定位于一年级经济学博士生的后继课程。本书有两大类读者。第一类读者运用本书作为微观经济计量学教材；一般来讲，在博士研究生的第二学年里讲授该课程，或者作为微观经济学领域的课程，比如以数据为导向的劳动经济学、公共经济学、行业组织等课程。第二类读者是研究者，将本书作为参考书，这些人虽然在微观经济计量学方面已经入门，但仍希望进一步提升自己在这方面的知识。

对于使用本书作为经济计量学教材的教师来说，一种最好的方式是，最初略过一些章节的方法，尽早引入基本的非线性横截面与线性面板数据模型。涉及重要方法的章节（第 5 章）涵盖了极大似然法与非线性最小二乘法估计。掌握极大似然法与非线性最小二乘法估计量的知识，为最广泛运用的非线性横截面模型（第 14

章、第 17 章和第 20 章)、基本线性面板数据模型(第 21 章)以及处理评估方法(第 23 章)提供了充足的基础知识。对于高级线性面板数据方法(第 22 章)来说,特别需要广义矩方法估计(第 6 章)知识。

对于利用本书作为参考书的读者来说,许多章节的写作尽可能地自成体系。值得注意的例外是,第 5 章与第 6 章偶尔出现某些一般估计结果的计算机命令,这是必需的。绝大多数章节模型的阐述采用使读者易于理解的讨论及例子作为开始。

www.econ.ucdavis.edu/faculty/cameron 网站给出了本书所用的全部数据与计算机程序,以及便于教学的相关有益材料。

写作本书是一项长期而艰巨的工作,有时显得遥遥无期。项目的完成得到了同事、朋友以及研究生的大力支持。我们特别感谢阅读并评论特定章节的下述人员:Bijan Borah, Kurt Brännäs, Pian Chen, Tim Cogley, Partha Deb, Massimiliano De Santis, David Drukker, Jeff Gill, Tue Gorgens, Shiferaw Gurmu, Lu JI, Oscar Jorda, Roger Koenker, Chenghui Li, Tong Li, Doug Miller, Murat Munkin, Jim Prieger, Ahmed Rahmen, Sunil Sapra, Haruki Seitani, Yachen Sun, Xiaoyong Zheng 和 David Zimmer。Pian Chen 对本书大部分内容给出了详细评论。我们感谢 Rajeev Ddhejia, Bronwyn Hall, Cathy Kling, Jeffrey Kling, Will Manning, Brian McCall 和 Jim Ziliak,他们为本书阐述许多实证例子提供了数据。我们感谢各个院系提供给作者的合作便利,以及为完成不同阶段的手稿而提供的便利。我们从两位匿名评阅人那里得到有益的评论。剑桥出版社的编辑 Scott Parris 为我们提供了极为宝贵的指导、建议和鼓励。

我们在经济计量学上的兴趣,源于我们在学生年代的学术生涯初期所受到的训练和所处环境的熏陶。第一位作者感谢澳大利亚国立大学;特别是斯坦福大学的 Takesh Amemiya 和 Tom MaCurdy;以及俄亥俄州立大学。第二位作者感谢伦敦经济学院和澳大利亚国立大学。

写作这样一本面向应用研究者的著作的兴趣,源自我们在各自学院(UC-Davis 与 IU-Bloomington)与研究生和同事进行研究时所显露的问题。

最后,要感谢我们的家庭,假如没有家庭的理解和支持,要想完成这本书是不可能的。

A. 科林·卡梅伦(A. Colin Cameron)　加利福尼亚州戴维斯

普拉温·K. 特里维迪(Pravin K. Trivedi)　印第安纳州伯明翰

目　录

CONTENTS

第五部分　面板数据模型

第六部分 深入专题

第一部分
预备知识

第一部分涵盖微观经济计量分析的基本内容——经济设定、统计模型以及数据集合。

第 1 章讨论微观经济计量学独具特色的方面，并且为本书提供一个概览。本章强调，数据的离散性、行为关系的非线性和异质性均是个体层面微观经济计量模型的重要方面。通过表述本书自始至终所采用的记号与习惯来结束本章。

第 2 章和第 3 章通过向读者介绍构成后面章节分析内容的重要模型与数据概念，为本书的剩余部分设置研究领域。

经济计量学的重要特点是，在各种不同统计复杂技巧层面上所描述的基本模型和数据概括，与囊括因果关系以及企图对因果参数加以估计的模型之间的重要差异。在经济计量学中，经典的因果性定义是通过由考尔斯委员会(Cowles Commission)联立方程模型捕获到的外生变量与内生变量之间以及结构式参数与简化式参数之间关键性的区别而得到的。尽管对于某些目的来说，简化式模型非常有用，但对于政策分析来说，结构参数或因果参数的知识极其重要。在联立方程框架下，对结构参数进行识别，产生了大量的概念上和实践上的困难。基于潜在结果模型的日益增多的可选择方法，同样试图识别因果参数，但是，这样做是在一种更易于处理的框架下以提出有限问题来实施的。第 2 章试图对在这些或其他可选择框架下产生的基本问题提供一种概述。最初发现这些挑战性材料的读者，在较好地熟悉本书后面内容所涵盖的特定模型之后，应该回到这一章上来。

实证研究者识别因果关系的能力不仅依赖于统计工具与模型，还依赖于可利用的数据类型。实验框架提供了一种建立因果联系的标准。然而，可观测的、非实验的数据形成了大多数经济计量推断的基础。第 3 章全面评述三种主要数据类型——可观测数据、源于社会实验的数据和源于自然实验的数据——的优缺点。对基于每一种数据类型所实施的因果推断的优缺点，也加以评论。

1 概 述

1.1 引 言

本书提供对微观经济计量分析(**mircoeconometric analysis**)的详细研究,即对个人或厂商的经济行为方面的个体水平数据进行分析。比较宽泛的定义还包括分组数据。通常,回归方法应用于横截面或面板数据。

对个体数据进行分析具有悠久的历史。厄恩斯特·恩格尔(Ernst Engel, 1857)是住户预算的最早的数量研究者。艾伦和鲍利(Allen and Bowley, 1935)、霍撒克(Houthakker, 1957)以及普雷斯和霍撒克(Prais and Houthakker, 1955)对随后同样的研究与建模传统做出重要贡献。在对微观经济计量学发展的激励中,同样具有影响的另外一些里程碑似的工作研究,包括在生产理论中由马歇克和安德鲁斯(Marschak and Andrews, 1944)所做的研究,以及在消费需求中由沃尔德和朱林(Wold and Jureen, 1953)、斯通(Stone, 1953)以及托宾(Tobin, 1958)所做出的那些研究。

与上面所提到的早期工作同样重要的是,关于住户预算和需求分析,本书中所涵盖的内容与离散选择分析、删失变量和截取变量模型的研究具有比较紧密的联系,在麦克法登(McFadden, 1973, 1984)和赫克曼(Heckman, 1974, 1979)的研究工作中,可以分别看到这些方面的第一个严谨的经济计量应用。这些研究并没有使用传统的线性模型,而早期研究极度信赖线性模型,并以此为特征。因此,它们曾导致经济计量学重要方法上的创新。马达拉(Maddala, 1983)和雨宫(Amemiya, 1985)的著作是研究这类内容(另外一些内容)的较早的教科书式的处理。正如赫克曼(Heckman, 2001)、麦克法登(McFadden, 2001)以及其他学者所强调的,建立在市场数据基础之上的处于支配地位的早期研究工作中的许多重要问题仍旧是重要的,尤其是关于因果经济关系可识别性的必要条件。然而,微观经济计量学的研究风格极多,多到足以写出一部完全致力于它的课本。

建立在个人水平、住户水平以及企业水平数据之上的现代微观经济计量学,拥有大量的出于横截面和纵向的样本调查,还有人口普查数据可以利用,这些数据都很容易获得。在过去 20 年里,随着个体水平上电子记录不断扩展和数据的收集,数据量呈现爆炸式增长。

同样，也可以利用计算机分析大量且复杂的数据集合。在许多情况下，可以利用事件水平数据；例如，市场营销学经常处理由超市电子扫描器所收集的买卖数据，而行业组织文献包括由在线订票系统收集的航空旅行数据的经济计量分析。现在，经济学存在一些新的分支，诸如社会实践和实验经济学，它们都会生成实验数据。这些发展创造了许多崭新的建模机会，而这种建模机会在仅仅利用汇总市场水平数据时是没有的。同时，数据量与类型爆炸式增长也产生了大量的方法问题。以揭示经济行为模式为目标，对这类大量微观数据进行加工处理与经济计量分析构成微观经济计量学的核心。对这类数据进行经济计量分析是本书的主题。

本书的重要先导内容是马达拉（Maddala，1983）与雨宫（Amemiya，1985）的书。像这两本书一样，本书涵盖了本科生和一年级研究生的经济计量学课程中的只是简要表述而非完全表述的专题。特别地，与雨宫（Amemiya，1985）的书相比，本书更以实践为导向。不过，在一些适当的地方，其阐述水平是高等的，对于数学推导弱于经济学学科的应用研究者来说，尤其是这样。

要求相对高等的阐述有几个原因。第一，数据常常是离散的或者删失的，在此情况下就要使用非线性方法（**nonlinear methods**），诸如 logit、probit 以及 Tobit 模型。这将导致建立在更困难的渐近理论基础之上的统计推断。

第二，对于这类数据来说，分布假设（**distribution assumptions**）是极为重要的。一种解答就是要充分发展详细地捕获到数据复杂性的高度参数模型，但是，这些模型面临估计的挑战。更为普遍的回答是要最小化参数假设，并且实施建立在标准误差基础上的统计推断，其中的标准误差对诸如异方差性和聚集（clustering）的复杂情况来说是“稳健的”。在这种情况下，尽管可以使用标准的回归软件包，但需要相当多的知识来确保有效的统计推断。

第三，经济研究通常的目的是要决定因果关系（**causation**），而不是仅仅测算相关关系，要采用观测数据而不是实验数据。这导致了脱离因果关系的一些方法，譬如工具变量、联立方程、测量误差相关、面板固定效应以及差异中的差分。

第四，一般来说，微观经济数据是利用横截面与面板调查、人口普查或者社会实验来收集。调查数据（**survey data**）则利用受限于复杂调查方法问题的这些方法来收集，违背简单随机抽样假设、样本选择问题、测量误差、不完全数据和/或缺失数据。对这类问题加以处理的方式支持从所估计的经济测量模型中得出的有效总体推断，这样做要使用高等方法。

最后，常常会有两个或更多的复杂情况同时发生，诸如具有面板数据的 logit 模型中的内生性。因此，详细阐述手册式的方法变得非常难以执行。相反，需要对支撑方法的理论进行相当深入的理解，这就如同研究者需要阅读经济计量学期刊文章以及应用标准的经济计量学软件一样。

1.2 微观经济计量学的特色

现在，我们考察微观经济计量学的几个优点，这由它的独有特性体现出来。

1.2.1 *离散性与非线性性*

首要的且最明显的特点是，微观经济计量数据通常是在低水平上汇总的。对于使用函数形式分析所关注的变量来说，这是一个重要的结果。在许多但不是最主要的情况下，可以证明，线性函数形式是不合适的。更为基本的是，非汇总会引起个人、厂商以及组织的最重要的异质性(**heterogeneity**)，如果人们要对基本关系做出有效推导，就应该对个人、厂商以及组织进行适当的控制(建模)。我们将在以下几节以较为详细的方式对这些问题加以讨论。

在微观数据中，汇总并不是全都没有，例如，当对家庭水平数据或企业水平数据进行汇总时，所汇集的总水平数量级经常比宏观分析中的普通情况要低一些。在宏观情况下，汇集过程会导致光滑，在求和过程中会令许多方向相反的运动相互抵消。汇总形成的变量常常表现出比其成分更为光滑的行为，而且汇总后变量的关系往往比其成分更具有光滑性。例如，在微观水平上，两个变量之间的关系可能是存在许多结点的分段线性关系。在汇总之后，这一关系可能由光滑函数很好地逼近。因此，不汇总的直接后果是，无论是变量本身还是变量之间的关系，都缺乏连续性与光滑性。

通常，个人水平与厂商水平数据涵盖大量的变异，不论是在横截面数据中，还是在时间序列数据中。例如，牛肉消费的每周平均值很可能是正的，而且光滑变动，然而，在给定的周里，个体家庭往往可以为零，而且有时可以转变为正值。由女工人提供的工作小时平均值不可能是零，但是，许多个体妇女具有零工作市场小时数(角点解)，而在她们的劳动市场历史过程中，在其他时间上却转变为正值。家庭假期开支平均值通常是正值，但在给定年份中，许多个体家庭假期开支具有零值。烟草制品人均消费通常是正值，但是总体中的许多个体从不消费这些产品，而且将永远不考虑价格与收入因素。如同帕德尼(Pudney, 1989)所发现的，微观数据表现出“遗漏、扭曲与角点”。遗漏对应于没有参加关注活动，扭曲对应于转换行为，而角点则对应于在特定时点上没有消费或没有参与。也就是说，响应的离散性与非线性是微观经济计量学所固有的。

微观经济计量学中，一类重要的非线性模型研究受限因变量(**limited dependent variables**)[马达拉(Maddala, 1983)]。这类模型包括许多模型，对于分析离散响应与带有受限变化范围的响应来说，提供了一种合适的框架。当然，对于分析宏观数据来说，如果需要的话，这种分析工具还是可以利用的。其要点是，它们是微观经济计量学中必不可少的，而且展示出其独特的性质。

1.2.2 *更加现实主义*

有时，宏观经济计量学(macroeconometrics)是建立在强假设基础之上的；代表性行为人假设就是一个重要事例。时常要求微观经济推理去判断实证结果的某些设定与解释是正确的。然而，几乎不可能说出，它们是如何以显性方式受到对时间与微观单元汇总的影响；否则，可做出非常极端的汇总假设。例如，汇总被说成是可以反映出假定的代表性行为的行为。

从微观经济理论观点来说，与建立在汇总数据上的那些数量分析相比，建立在微观数据上的数据分析，被认为更具有现实性。判断这一陈述正确有三个理由。第一，在这种假设下，所设计的变量测量常常更为直接(尽管不一定无测量误差)，而且更对应所要检查的理论。第二，关于经济行为的假设，通常是从个体行为理论中发展出来的，如果这些假设是利用汇集数据进行检验，那么就可以做出许多近似和简化假设。这种代表性行为的简化假设导致信息大量损失，且严重地限制了实证研究的范围。由于微观经济计量学可以避免这类假设，而且原则上常常如此，所以微观数据提供了更加现实的用于检验微观经济假设的框架。这并不是声称一定要在实证研究工作中得到微观数据。需对这种陈述进行逐一判断。最后，经济活动的现实描绘应该提供作为个体异质性后果的广泛结果及响应，并且可以通过基本理论进行预测。在这个意义上，微观经济数据集能支持更现实的模型。

微观经济计量数据经常是从住户或厂商调查中得到的，一般包含广泛的行为，其中的许多行为结果采用离散或分类形式。这种数据集具有许多难以处理的特性，要求在用公式表示和分析它们时使用特殊工具，虽然宏观经济计量研究中并不完全缺乏这种情况，然而，特殊工具仍然没有得到广泛使用。

1.2.3 *更多信息内容*

如果微观数据集具有信息价值，那么微观数据集的潜在优点就可以实现。由于样本调查经常提供成千上万个横截面单元的独立观测值，而通常为高度序列相关的标准宏观时间序列一般至多由几百个观测值构成，与之相比，前者更具有信息价值。

如同下一章将要解释的，由于微观数据可能具有相当大的噪声，实践中的情况并不能如此丢掉。在个体层面上，许多(特质的)因素在决定响应时起着很大的作用。这些因素常常是不能观测到的，导致人们在随机成分标题下对它们加以处理，具有相当大的观测变异部分。在此意义上，随机性在微观数据中具有更大的作用。当然，这会影响到回归的拟合优度测算。最初，通过汇总时间序列分析探索经济计量学的大学生，经常以看到大的 R^2 值为条件，当初次遇到横截面回归时，他们对回归方程的“低解释能力”会表现出失望或惊讶。然而，存在着强假设，即至少在某个范围内，很大的微观数据集具有很高的信息价值。

另外一个限制条件是，当人们研究纯横截面数据时，几乎很少能对所研究的跨时关系方面讲些什么。这种特殊行为能够利用面板数据和过渡数据加以研究。

在特殊情况下，人们在某种特定的经济环境下，对特定经济行为人的群体行为响应感兴趣。一个事例是失业保险对青年失业者的工作搜寻行为的影响。另一个事例是，接受收入保障金的低收入个体对劳动力供给的响应。除非使用微观数据，否则这类问题在实证工作中不能直接加以讨论。

1.2.4 *微观经济基础*

在不同的经济理论中，经济计量模型所起的作用是不同的。在有模型的情况下，先前的理论在对模型进行设定和选择估计方法时起着支配作用。而在另一种

实证研究情况下，却很少使用经济理论。

在第一种情况下，分析的目的是要识别和估计那种刻画个体的口味与偏好以及/或技术关系的基本参数，有时称之为深参数。作为一种简称，我们将这一方法称为结构方法(**structural approach**)。其特点紧密地依赖于经济理论，并强调因果推断。这种模型需要许多假设，例如，对成本函数或生产函数进行准确设定，或对误差项分布进行设定。运用这种方法的实证结论在背离假设的情况下是不稳健的。在 2.4.4 节，我们会更多地谈及此方法。现在，我们直接强调，如果结构方法可利用汇总数据加以实施，它只在非常严格的(而且可能不现实的)条件下得到基本参数的估计值。微观数据集为结构方法提供了更有前途的环境，因为在模型设定中，本质上允许它们更具有灵活性。

在第二种情况下，分析的目的是在变量由研究者给定或为外生变量的条件下，对所关注的响应变量之间的关系进行建模。内生性(**endogeneity**)或外生性(**exogeneity**)的更正式定义将在第 2 章给出。作为一种简记名称，我们将这一方法称为简化式方法(**reduced form approach**)。其基本思想是，简化式分析并不总是考虑所有因果的相互依存。这种回归模型关注于给定回归元 $\mathbf{x}$ 对 y 的预测，而不是关注于回归参数的因果解释，这常常称为简化式回归。如同在第 2 章将要阐述的，简化式模型的参数通常是结构参数的函数。如果没有结构参数的某种信息，它们就不是可解释的。

1.2.5 *非汇总与异质性*

有时，据说宏观经济计量学的问题和争论来自宏观时间序列的序列相关，而微观经济计量学的问题和争论则出自个体水平数据的异质性。在许多微观经济计量分析中，尽管这是对努力建模的良好刻画，但是它需要加强，并且受限于重要的限制条件。在微观经济计量模型中，对动态相依性进行建模或许是一个重要的问题。

非汇总的好处已在本节前面强调过，但是，它要付出代价：因为数据越是非汇总的，则对个体问题的异质性加以控制就越发重要。异质性，或者更准确地讲为不可观测的异质性，在微观经济计量学中起着非常重要的作用。显然，反映个体间异质性的许多变量，诸如性别、民族、教育背景、社会以及人口因素都是直接可观测的，从而对它们能够加以控制。与之相比，个体的动机、能力、智力等方面的差异或者是不可观测的，或者充其量也不过是不完全可观测的。

最简单的反应就是忽略这种异质性，即将其并入回归扰动之中。这毕竟是人们如何处置无数多个很小的不可观测因素的方法。当然，这一步会增加变异的未解释部分。更严重的是，一旦忽视持久的个体差异，将会导致与其他作为持久的个体间差异来源的因素相混淆(**confounding**)。在关注的变量中，对差异而言，如果不同的回归元(预测元变量)的个体贡献不能在统计形式上得以分离，就会发生混淆。例如，假定把因素 x_1(受教育)说成是 y(收入)的变异来源，而另一个变量 x_2(能力)作为变异的其他来源并没有出现在模型中。于是，总变异中归因于第二个变量的那一部分，被错误地归因于第一个变量。从直观上讲，它们的重要性被混淆了。混

淆偏倚的重要来源是从模型中不正确地省略回归元，并且包括代表省略变量的其他变量。

例如，考察下述情况，在带有回归元向量 $\mathbf{x}$ 的回归均值函数中，包括项目参与（0/1 虚拟）变量 D：

$$y=\mathbf{x}'\boldsymbol{\beta}+\alpha D+u \tag{1.1}$$

其中，u 表示误差项。“处理”（treatment，又称干预）这一术语用于生物或实验科学之中，表示涉及某实验中参与者的实施方式。在经济计量学中，它通常是指，参与某一项可以影响到关注结果的活动。这项活动可以随机指派参与者，也可以由参与者自己选择。因此，尽管众所周知，个人选择其自己的受教育年数，但人们还是将受教育年数看成是“处理”变量。假定项目参与被设为离散变量。“处理变量”的系数 α 测算以协变量为条件的项目参与（$D=1$）的平均影响，如果人们不控制不可观测的异质性，则潜在的模棱两可性会影响到对结果的解释。如果发现 D 具有显著影响，那么会产生下述问题：因为 D 与某些影响 y 的不可观测变量是相关的，或者在 D 与 y 之间存在因果关系，那么 α 会显著地异于 0 吗？例如，当所考虑的项目是大学教育，并且协变量不包括能力的测量，要给出一种完全的因果解释就会受到质疑。因为这一争论很重要，所以应该更多地关注如何控制异质性。

在一些涉及动态考虑的情况下，可利用的数据类型或许会限制人们如何控制异质性。考察两个住户，除了其中一个表现出更偏好于消费商品 A 之外，其余的方面都一样。人们可以通过允许个人效用函数包含反映他们不同偏好的异质性参数来对此加以控制。现在，假定存在一种消费者行为理论，它声称消费者偏好商品 A，在此意义下，消费者在一个时期里消费它越多，则消费者在未来消费它也越多的可能性就越大。这种理论提供了对商品 A 消费的持久个体间差异的另外一种解释。一旦对异质性偏好加以控制，检验消费中哪一种持久性来源——异质性偏好或嗜好——可以解释各种不同消费模式就成为可能。每当在可观测的结果中某个动态元素产生了持久性，便出现这类问题。这类问题的几个事例在本书中的不同地方都出现过。

对异质性进行建模的一系列方法，在微观经济计量学中同时并存。对其中的一些将会简要提及，而详细内容则推迟到后面阐述。

一种极端的求解是忽略掉全部不可观测的个体间差异。如果不可观测的异质性与可观测的异质性是不相关的，并且如果所研究的结果没有时期间的相依性，就不会产生上述问题。当然，存在一些强假设，并且甚至满足这些假设，不是所有的经济计量困难都会消失。

处理异质性的一种方法是把它看作固定效应（**fixed effect**），并把它估计成个体特定 0/1 虚拟变量的系数。例如，在横截面回归中，允许每一个微观单元拥有自己的虚拟变量（截距）。因为当样本添加一个新个体时，也会增加一个新的截距参数，所以这会造成参数急剧增多，因此，如果我们的数据都是横截面的，那么这一方法将不起作用。当可以利用每一个个体单元的多重观测值时，最普遍的形式是对 N 个横截面单元的每一个都具有 T 个时间序列观测值的面板数据，例如，如果模型

是线性的，且固定效应是可加的，通过一阶差分来估计固定效应或者剔除固定效应就是可行的。如果模型是非线性的，并且固定效应通常不是可加的，就需要考虑其他的一些方法。

对不可观测异质性进行建模的第二种方法是，通过随机效应（**random effects**）模型来建模。随机效应模型拥有许多不同的公式表述方法。一种流行的公式是假定一个或多个回归参数，常常是回归截距，对于不同的横截面会随机变化。在另一种公式中，回归误差是给定的分量结构含有特定个体的随机成分。然后，随机效应模型企图从那种推导出的随机成分中去估计分布参数。在一些情况下，比如需求分析，随机项可以被解释成为随机偏好变异。随机效应模型利用横截面数据或面板数据得到估计。

1.2.6 *动态特性*

在横截面分析中，一个非常普遍的假设是没有跨时期相依性，即缺少动态学。因此，它隐含地假定，观测值对应于随机均衡，而对均衡的偏离则由序列独立随机扰动来表示。甚至在微观经济计量学中，对于某种数据情况来说，这样的假设或许太强了。例如，它与序列相关不可观测异质性的存在不一致。有关滞后因变量的相依性也会违背此假设。

上述讨论已经阐述单一横截面分析的一些潜在局限性。如果可以利用重复的横截面数据，就可克服一些局限性。然而，如果利用重复的横截面数据，那么引起最少争议的方法应该是使用面板数据为好。

1.3 全书概览

全书分成六大部分。第一部分阐述微观经济计量建模所涉及的问题。第二部分和第三部分阐述非线性回归模型的估计及统计推断的一般性理论。第四部分和第五部分分别专门研究应用微观经济计量学中使用的核心模型，这里既有横截面模型，又有面板数据模型。第六部分涵盖比较广泛的专题，大部分要利用前面一些章节所阐述的内容。

本书内容概览如表 1.1 所示。本节将依次详述每个部分。

1.3.1 *第一部分：预备知识*

第 2 章和第 3 章对微观经济计量（**microeconometric**）的建模方法特性，以及处于更一般统计领域回归分析中的微观经济数据结构（**microeconomic data structures**）展开讨论。全书自始至终地对这两章中所出现的许多问题进行不断深入探讨，以便为读者开发一些必要的工具。

1.3.2 *第二部分：核心方法*

第 4 章～第 10 章详述经典估计和统计推断中使用的重要的一般方法。特别是，第 5 章给出的结果广泛地应用于全书。

表 1.1 全书概览

部分及章节	背景[a]	事 例
一、预备知识		
1. 概述	—	
2. 因果模型与非因果模型	—	联立方程模型
3. 微观经济数据结构	—	观测数据
二、核心方法		
4. 线性模型	—	普通最小二乘法
5. 极大似然法与非线性最小二乘法估计	—	m 估计或极值估计
6. 广义矩方法与系统估计	5	工具变量
7. 假设检验	5	沃尔德得分以及似然比检验
8. 设定检验与模型选择	5、7	条件矩检验
9. 半参数方法	—	核回归
10. 数值最优化	5	牛顿—拉夫森迭代法
三、基于模拟的方法		
11. 自助法	7	百分位数 t-方法
12. 基于模拟的方法	5	极大模拟似然
13. 贝叶斯方法	—	马尔可夫链蒙特卡罗
四、横截面数据模型		
14. 二值结果模型	5	logit，关于 $y=(0, 1)$的 probit
15. 多项式模型	5、14	关于 $y=(1, \cdots, m)$的多项式 logit
16. Tobit 模型与选择模型	5、14	关于 $y=\max(y^*, 0)$的 Tobit
17. 过渡数据：生存分析	5	关于 $y=\min(y^*, c)$的考克斯比例风险
18. 混合模型与不可观测异质性	5、17	不可观测异质性
19. 多重风险模型	5、17	多重风险
20. 计数数据模型	5	$y=0,1,2,\cdots$泊松模型
五、面板数据模型		
21. 线性面板模型：基础	—	固定效应与随机效应
22. 线性面板模型：扩展	6、21	动态与内生回归元
23. 非线性面板模型	5、6、21、22	面板 logit、Tobit 以及泊松模型
六、深入专题		
24. 分层样本与整群样本	5	对于不同的 j，数据$(y_{ij}, \mathbf{x}_{ij})$相关
25. 处理评估	5、21	如果参与项目，回归元 $d=1$
26. 测量误差模型	5	带有观测误差的 logit 模型
27. 缺失数据与估算	5	带有缺失观测值的回归

[a]背景给出了除第 4 章普通最小二乘法和加权最小二乘法处理之外所需要的基础章节。注意，第一个面板数据章(第 21 章)仅要求第 4 章的内容。

第 4 章阐述线性回归模型(**linear regression model**)的一些结果，强调与本书其余内容最相关的那些问题和方法。相对来说，这种分析直接且简单，因为线性模型估计量存在着显性表达式，譬如普通最小二乘法。

第 5 章和第 6 章阐述能用于估计量通常没有显性解的那些非线性模型上的估计理论(**estimation theory**)。渐近理论可用于获得估计量的分布,着重于获得依赖相对弱分布假设的稳健误差估计值。第 5 章阐述相当一般的估计处理和专门化的非线性最小二乘法,以及极大似然估计。第 6 章分别给出更富有挑战性的广义矩方法估计量以及专门化的工具变量估计。

第 7 章阐述当估计量关于参数是非线性的以及所要检验的假设关于参数可能是非线性的时候,对经典假设进行检验(**classical hypothesis testing**)。设定检验(**specification tests**)和假设检验是第 8 章的主题。

第 9 章阐述半参数估计(**semiparametric estimation**)方法,譬如核回归。重要的事例是对条件均值的灵活建模。对于专利事例而言,非参数回归模型是 $E[y|x]=g(x)$,其中,函数 $g(\cdot)$ 表示未设定的,并且要用估计值来代替。于是,估计具有无限维成分 $g(\cdot)$,从而导致了非标准渐近理论。就另外的回归元来说,这需要某种进一步的结构,称这种方法为半参数的或者半非参数的。

第 10 章阐述当估计量是以隐性方式定义的,经常作为某些一阶条件的解时,可用于计算参数估计值的计算方法(**computational methods**)。

1.3.3 *第三部分: 基于模拟的方法*

第 11 章～第 12 章考察依赖模拟的估计和推断的方法。这些方法通常更为密集计算,与第二部分所阐述的方法相比,这些方法当前很少被应用。

第 11 章阐述用于统计推断的自助法[1](**bootstrap method**)。借助于模拟获得新样本,例如,通过从最初样本中重复进行放回再抽样,就会产生估计量的经验分布。当从渐近理论得出的公式很复杂时,自助法能提供一种简单获得标准误差的方法,就如同某种两步估计量的情况。进一步地,如果实施恰当,那么自助法会导致小样本的统计推断。

第 12 章阐述基于模拟估计方法(**simulation-based estimation methods**),这会涉及对不存在闭型解的概率分布进行积分的那类模型。通过从有关分布与平均所做出的多重推断来进行估计还是可能的。

第 13 章阐述贝叶斯方法(**Bayesian methods**),把观测到的数据分布与参数的设定先验分布结合起来,获得作为估计基础的参数的后验分布。尽管后验分布不存在闭型解,但最近进展使得计算其解成为可能。贝叶斯分析能够提供与经典方法大不相同的用于估计与推断的方法。然而,在许多情况下,只有贝叶斯工具箱才能对其他方法难以解决的问题进行估计和推断。

1.3.4 *第四部分: 横截面数据模型*

第 14 章～第 20 章阐述横截面数据(**cross-section data**)的主要非线性模型。这一部分是本书的核心,同时阐述一些高等专题,譬如受限因变量模型与样本选择。这类模型是通过因变量取值范围来定义的。

[1] 又称自举、再抽样。——译者注

第 14 章提供了因变量仅仅能取两个可能值——譬如取 $y=0$ 或 $y=1$——的二值数据(**binary data**)模型。第 15 章阐述对因变量取几个离散值的多项式(**multinomial**)模型,这是对上一章内容的扩展。一些事例包括就业状况(就业、失业和非劳动力)以及上下班所选的交通方式(小汽车、公交车或火车)。线性模型可以提供信息,却并不恰当,因为线性模型能产生单位区间以外的预测概率值。相反,要使用 logit、probit 以及其他有关的模型。

第 16 章阐述带有删失(**censoring**)、截取(**truncation**)、样本选择(**sample selection**)的一些模型。一些事例包括以选择工作为条件的年度工作小时数,还有以住院治疗为条件的医院开销。在这些情况下,数据在 $y=0$ 时是一组不完全观测到的观测值,而且在 $y>0$ 时仍旧如此。可以证明,即使基本过程是线性的,但这样观测到数据的模型是非线性的,而且关于观测到数据的线性模型可以使人严重产生误解。对删失、截取或者样本选择的简单纠正,比如 Tobit 模型,都是存在的,但是,这些却非常依赖于分布假设。

持续期限数据(**duration data**)模型将在第 17 章～第 19 章加以阐述。其中一个事例就是失业时段的长短。标准回归模型包括指数模型、威布模型以及考克斯比例风险模型。附带说一句,如同第 16 章一样,因变量经常是不完全观测到的。例如,处于当前时段长度的数据就是不完全的,而不是完全时段的长度。

第 20 章阐述计数数据(**count data**)模型。一些事例包括对健康的各种测量,例如,医生出诊次数以及住院天数。该模型又是非线性的,因为条件均值都是非负的。重要的参数模型包括泊松模型和负二项式模型。

1.3.5 *第五部分: 面板数据模型*

第 21 章～第 23 章阐述面板数据(**panel data**)方法。这里,对于样本中众多个体的每一个而言,一系列时期的数据都是可观测的,因此,因变量与回归元既要标记个体,又要标记时间。对于给定个体来说,任何分析都需要控制不同时期中误差项的可能正相关。附带讲一句,面板数据能提供充足的数据来控制不可观测的特定个体的常值效应,与仅可利用横截面时所需要的那些假设相比,在更弱的假设下,识别因果关系是可能的。

第 21 章阐述基本线性面板数据模型,着重于固定效应(**fixed effects**)模型和随机效应(**random effects**)模型。第 22 章阐述允许滞后因变量与内生回归元对线性模型的扩展。第 23 章阐述有关第四部分中非线性模型的面板方法。

面板数据方法放在本书后面,这是为了提供一种统一独立自足式的处理。第 21 章本可以放在第 4 章后面,只要有最小二乘法估计的内容,就可以将其阐述得通俗易懂。

1.3.6 *第六部分: 深入专题*

这一部分考察与第四部分和第五部分通常有关的全部模型的重要专题。第 24 章研究对几种不同模型中整群数据的建模。第 25 章讨论处理评估。处理评估是一个一般性术语,它涵盖一系列关注于某种“处理”影响测量的广泛的模型,其中

的处理不是以外生方式就是以随机方式指派给某个关注测量的个体的，记为“结果变量”。第 26 章研究结果变量和/或回归元变量上的测量误差的后果，着重于某些重要的非线性模型。第 27 章考察某些处置线性和非线性回归模型中缺失数据的方法。

1.4 如何使用本书

本书假定读者已具备对矩阵代数线性回归模型的基本认识。与格林(Greene，2003)的书相比，本书写作定位于博士生一年级后续课程的数学水准上。

尽管本书的一些内容已涵盖一年级后续课程，但其大部分内容看起来像经济学博士生二年级课程，或者是以数据为导向的微观经济学领域课程，譬如劳动经济学、公共经济学或者行业组织学。本书既可用于经济计量学教材，又可用于此领域课程的补充读物。更一般地讲，本书目的旨在作为经济学、有关的社会科学——比如社会学、政治科学以及流行病学——领域中应用研究者的有益参考书。

对于利用本书作为参考书的读者来说，许多模型章节尽可能写成独立式的。对于第四部分和第五部分所阐述的特定模型而言，除了必须掌握第 5 章和第 6 章中在某些情况下的一般估计结果之外，以独立方式阅读有关章节通常就足够了。相当多的章节是从易于广大读者理解的讨论和例题开始的。

对于利用本书作为课程教材的教员来说，最好是尽可能早地略过许多方法性章节，引进基本的非线性横截面或线性面板数据。第 14 章～第 16 章阐述最普遍使用的非线性横截面模型，这些都需要第 5 章所阐述的极大似然与最小二乘法估计。第 21 章的线性面板数据模型甚至需要更少的预备知识，基本上只需第 4 章的知识。

表 1.2 提供了在加利福尼亚大学戴维斯为二年级研究生做半学期课程教学的提纲。一个学期可提供足够多的时间涵盖这个提纲前面一半章节中给出的基本结果。如果还有时间，人们能够更进一步深入研究涵盖第 11 章～第 13 章一部分密集计算的估计方法(基于模拟的估计、自助法，这曾在第 7 章简略地阐述过，还有贝叶斯方法)；另外，第 17 章～第 20 章阐述横截面模型(持续期限和计数)；此外，第 22 章和第 23 章给出面板数据模型(线性模型的扩展以及非线性模型)。

表 1.2 10 周 20 次讲座提纲

讲座	章	专 题
1～3	4、附录 A	线性模型和渐近理论回顾
4～7	5	估计：m 估计、ML 与 NLS
8	10	估计：数值最优化
9～11	14、15	模型：二值与多项式
12～14	16	模型：删失与截取
15	6	估计：GMM
16	7	检验：假设检验
17～19	21	模型：基本线性面板
20	9	估计：半参数

在印第安纳大学伯明翰，以15周为一学期的微观经济计量学课程，建立在第四部分和第五部分大多数内容的基础上。就此课程而言，其必备条件课程所涵盖的内容和第二部分中的相似。

在前三章的导论之后，每一章的结尾都提供一些练习题。这些练习是边学边练的性质：一些习题纯粹是关于方法上的，而另外一些习题则需要对生成数据或实际数据进行分析。问题的困难程度大部分与专题的困难程度有关。

1.5 软件

存在众多用于数据分析的软件包。流行的、强有力的微观经济计量软件包有LIMDEP、SAS以及STATA，所有这些软件都提供一种关于预先编制好的程序的广泛范围，而且支持用户利用矩阵编程语言定义程序。另外，一些同样被广泛使用的软件包括EVIEWS、PCGIVE以及TSP。尽管这些软件均是面向时间序列的，但它们能够支持某种横截面数据分析。那些希望自己编程的用户还可利用一系列可供选择的软件，包括GAUSS、MATLAB、OX以及SAS/IML。有关这些软件包的最新详细信息以及许多其他软件包，能够有效地通过互联网浏览器和搜索引擎来准确地寻找到。

1.6 记号与习惯

本书广泛地使用向量与矩阵代数。

向量被定义为列向量，并用小写黑体字母表示。例如，对于线性回归而言，回归元向量 $\mathbf{x}$ 表示 $K\times1$ 维列向量，其第 j 个元素为 x_j，而参数向量 $\boldsymbol{\beta}$ 表示列向量，其第 j 个元素为 β_j，因而有：

$$\underset{(K\times1)}{\mathbf{x}}=\begin{bmatrix}x_1\\ \vdots\\ x_K\end{bmatrix}\quad 且 \quad \underset{(K\times1)}{\boldsymbol{\beta}}=\begin{bmatrix}\beta_1\\ \vdots\\ \beta_K\end{bmatrix}$$

于是，线性回归模型 $y=\beta_1x_1+\beta_2x_2+\cdots+\beta_Kx_K+u$ 可表示为 $y=\mathbf{x}'\boldsymbol{\beta}+u$。有时候，第 i 个观测值要添加下标 i。于是，第 i 个观测值的线性回归方程是：

$$y_i=\mathbf{x}_i'\boldsymbol{\beta}_i+u_i$$

样本是 N 个观测值 $\{(y_i,\ \mathbf{x}_i),\ i=1,\cdots,\ N\}$ 的其中之一。在本书中，通常假定观测值对于不同的 i 是独立的。

矩阵利用大写黑体字母来表示。在矩阵记号中，样本表示成 $(\mathbf{y},\ \mathbf{X})$，其中，$\mathbf{y}$ 表示 $N\times1$ 维向量，其第 i 个元素为 y_i，而 $\mathbf{X}$ 表示第 i 行为 $\mathbf{x}_i'$ 的矩阵，因而有：

$$\underset{(N\times1)}{\mathbf{y}}=\begin{bmatrix}y_1\\ \vdots\\ y_N\end{bmatrix}\quad 且 \quad \underset{(N\times\dim(\mathbf{x}))}{\mathbf{X}}=\begin{bmatrix}\mathbf{x}_1'\\ \vdots\\ \mathbf{x}_N'\end{bmatrix}$$

于是，一旦将所有 N 个观测值叠放在一起，则线性回归模型是：

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{u}$$

其中，$\mathbf{u}$ 表示 $N\times 1$ 维列向量，第 i 个元素为 u_i。

矩阵记号虽简洁，但有时把矩阵的乘积写成向量乘积之和更为清楚。例如，OLS 估计量等价地写成下述两种方式之一：

$$\hat{\boldsymbol{\beta}}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}=\left(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sum_{i=1}^{N}\mathbf{x}_i y_i$$

普通的参数记号表示成 $q\times 1$ 维向量。回归参数则利用 $K\times 1$ 维向量 $\boldsymbol{\beta}$ 来表示，它可以等于 $\boldsymbol{\theta}$，或者是 $\boldsymbol{\theta}$ 的子集，这要依赖于背景而定。

本书使用许多缩写符号和首子母缩略词。表 1.3 总结出某些常用估计方法所使用的缩写符号，它们依照估计量是阐明线性回归模型还是非线性回归模型来排序。我们还使用下述缩写方式：dgp（数据生成过程）、iid（独立同分布）、pdf（概率密度函数）、cdf（累积分布函数）、L（似然）、$\ln L$（对数似然）、FE（固定效应）以及 RE（随机效应）。

表 1.3　常用首字母缩略词和缩写方式

线性	OLS	普通最小二乘法
	GLS	广义最小二乘法
	FGLS	可行广义最小二乘法
	IV	工具变量
	2SLS	两阶段最小二乘法
	3SLS	三阶段最小二乘法
非线性	NLS	非线性最小二乘法
	FGNLS	可行广义非线性最小二乘法
	NIV	非线性工具变量
	NL2SLS	非线性两阶段最小二乘法
	NL3SLS	非线性三阶段最小二乘法
普通	LS	最小二乘法
	ML	极大似然法
	QML	准极大似然法
	GMM	广义矩方法
	GEE	广义估计方程

2 因果模型与非因果模型

2.1 引 论

微观经济计量学是研究由关于个体、家庭以及厂商的微观数据所发展起来的数据分析方法的理论及其应用。较为宽泛的定义还可包括地区数据和州数据。微观数据通常或者是横截面的(数据涉及在同一时点上的一些状况),或者是纵向的(面板的)(数据涉及一些历经几个时期的相同的观测单元)。这种观测值既可以由非实验方案生成,譬如人口普查和调查,又可以由准实验或实验方案生成,譬如由政府实施的自愿者参与实验。

微观经济计量模型可以是对一系列微观经济观测值的概率分布的一个完全设定;也可以是对一些变量的某种分布性质的部分设定,譬如矩。特别是,关注以回归元为条件的单个因变量的均值。

微观经济计量学有几个目的。这些目的既包括数据描述,又包括因果推断。第一种情况可以被广泛地定义,以便于包括响应变量的矩性质,或者强调关联而不是因果关系的回归方程。第二种情况包括因果关系,其目的在于对微观经济行为进行测量,或对微观经济行为的假说与命题的实证进行证实或反驳。因此,实证研究的类型和方式可以有很广泛的范围。其中一种极端情况是以高度结构化建立起来的结构模型,它是由对基本经济行为的详细设定推导出来的,用以分析相互依存的微观经济变量的因果(**causal**)行为或者结构关系(**structural relationships**);另外一种极端情况是简化式(**reduced form**),它没有必要依赖于对所有相关的相互依存变量进行详细的设定,目的是研究变量之间的相关性与关联性。两种方法都分享有助于理解微观经济行为的揭示重要而引人注目关系的共同目标,但是,在指导其实证研究的过程中,它们依赖于经济理论的程度是有差异的。

作为一门学科分支的微观经济计量学,比起关注于对市场和汇总数据建模的宏观经济计量学要“年轻”一些。应用经济计量学的早期大量研究是建立在由政府机构收集的总时间序列的基础之上。有关统计需求分析的早期大多数研究,在20世纪40年代以前一直使用的是市场数据,而不是个体或家庭数据[亨德里和摩根(Hendry and Morgan, 1996)]。摩根(Morgan, 1990)关于经济计量思想历史的书,除了有一个重要的例外,几乎没有涉及20世纪40年代以前的微观经济计量学

著作。那个例外是关于家庭收支预算数据的研究，即对许多家庭不富裕的生活标准进行的研究。这曾导致对家庭收支预算数据的搜集，从而为某些早期的微观经济计量研究，比如艾伦和鲍利(Allen and Bowley, 1935)那些先驱性的研究，提供了原始资料。然而，只有在 20 世纪 50 年代，微观经济计量学才作为一个独特的有组织的学科分支而出现。甚至进入 20 世纪 60 年代，微观经济计量学的核心是由建立在家庭调查基础之上的需求分析构成的。

随着 2000 年度诺贝尔经济学奖授予詹姆斯・赫克曼(James Heckman)和丹尼尔・麦克法登(Daniel McFadden)，以表彰他们对微观经济计量学的贡献，该学科领域作为一个独特的学科分支才形成了清晰的框架。这个奖项表彰赫克曼"在分析选择样本的理论和方法上的发展"，同时表彰麦克法登"在分析离散选择的理论和方法上的发展"。在提及微观经济学所研究论题类型的事例时，人们会引经据典："……是什么因素会促使个体决定是否去工作？要是工作的话，做多少小时呢？经济激励是怎样影响个体对教育、职业或住所地点的选择的？各种不同的劳动力市场与教育计划对个体收入和就业的影响是什么呢？"

微观经济计量学方法的应用不仅出现在微观经济学的每一个领域中，而且出现在其他同类的社会科学之中，比如政治科学、社会学以及地理学。

从 20 世纪 70 年代开始，而且特别是在过去 20 年间，我们处理大量数据集以及有关的计算能力都发生了革命性的进步。这些连同大量可利用性微观经济数据集的剧增都极大地扩展了微观经济计量学的范围。因此，尽管实证需求分析继续成为微观经济计量方法应用的最重要领域之一，但是它们的类型与内容严重地受到崭新方法与模型的影响。更进一步地，经济发展、金融、健康、工业组织、劳动力和公共经济学、应用微观经济学的应用现在成为平常之事，而且这些应用在本书的不同地方都将遇到。

本书主要关注过去 30 年间所出现的较新内容。我们的目标是研究一些概念、模型以及方法，我们认为，这些内容是现代微观经济计量学家工具箱中的标准组成部分。当然，作为本书假定的读者与作者自己背景的函数，一些标准方法与模型的概念必定既有主观性又有弹性。我们认为，还会存在相对于引论书而言更高等的一些论题，譬如以不同范畴设置的其他情况。

微观经济计量学关注于能给出结构解释的非线性模型的含义以及可获得的估计值。本书的大部分内容，特别是第二部分至第四部分，将阐述非线性模型的方法。这些非线性方法和包括生物统计学的许多应用统计学领域相交叉。与之相比，经济计量学的显著特性是对因果建模的强调。本章引进和因果(以及非因果)建模相关的重要概念，以及既与线性模型关系联系密切又与非线性模型关系联系密切的概念。

第 2.2 节和 2.3 节引入重要的结构与外生性的概念。第 2.4 节使用线性联立方程组模型作为结构模型的特殊解释，而且将它与其他的重要简化式模型的概念联系起来。第 2.5 节给出识别的定义。第 2.6 节考察单方程结构模型。第 2.7 节则引入潜在结果模型，并且把潜在结果模型中的因果参数和解释与联立方程中的那些内容加以比较。第 2.8 节对建模和估计策略给出一个简要的讨论，以便应对

计算和数据的挑战。

2.2 结构模型

结构(**structure**)具有以下四个性质：

1. 为了方便起见，把一系列变量 **W**(“数据”)分割成 [**Y Z**]；

2. **W** 的联合概率分布为 $F(\mathbf{W})$；

3. 依照假设的因果与效应关系；对 **W** 定出先验顺序，并对已假定模型的先验约束进行设定；

4. 对函数形式设定参数、半参数或者非参数的形式，并且对模型的参数加以约束。

这种结构模型的一般描述和已为大家所接受的考尔斯委员会(Cowles Commission)对结构的定义一致。例如，萨根(Sargan, 1988, 第 27 页)写道：

> 模型是关于一系列观测值的概率分布的设定。结构是对那个分布参数的设定。因此，结构是对所有参数都指定数值的模型。

我们考察下述情况，即建模目的是解释可观测向量值变量 $\mathbf{y}$, $\mathbf{y}'=(y_1,\cdots,y_G)$。$\mathbf{y}$ 的每一个元素都是 $\mathbf{y}$ 的某些其他元素与解释变量 $\mathbf{z}$ 以及一个纯随机扰动项 u 的函数。注意到，假定变量 $\mathbf{y}$ 是相互依存的。与之相比，不可以对 $\mathbf{z}_i$ 之间的相互依存进行建模。第 i 个观测值满足隐性方程集合：

$$\mathbf{g}(\mathbf{y}_i,\ \mathbf{z}_i,\ \mathbf{u}_i \mid \boldsymbol{\theta})=\mathbf{0} \tag{2.1}$$

其中，$\mathbf{g}$ 表示一个已知函数。我们把这称为结构模型(**structural model**)，把 $\boldsymbol{\theta}$ 称为结构参数。这对应于本节前面曾给定的性质 4。

假定对于每一个$(\mathbf{z}_i,\ \mathbf{u}_i)$而言，关于 $\mathbf{y}_i$ 存在唯一解。于是，我们能够以显性形式把关于 $\mathbf{y}$ 的方程写成$(\mathbf{z},\ \mathbf{u})$的函数：

$$\mathbf{y}_i=\mathbf{f}(\mathbf{z}_i,\ \mathbf{u}_i \mid \boldsymbol{\pi}) \tag{2.2}$$

这称为结构模型的简化式(**reduced form**)，其中，$\boldsymbol{\pi}$ 表示简化式参数向量，它是 $\boldsymbol{\theta}$ 的函数。该简化式可通过求解给定$(\mathbf{z}_i,\ \mathbf{u}_i)$时关于内生变量 $\mathbf{y}_i$ 的结构模型来获得。简化式参数 $\boldsymbol{\pi}$ 是 $\boldsymbol{\theta}$ 的函数。

如果建模目的是对 $\boldsymbol{\theta}$ 的元素进行推断，那么式(2.1)提供了推断的直接途径。这涉及对结构模型的估计。然而，由于 $\boldsymbol{\pi}$ 的元素都是 $\boldsymbol{\theta}$ 的函数，所以式(2.2)还提供了对 $\boldsymbol{\theta}$ 推断的间接途径。如果 $\mathbf{f}(\mathbf{z}_i,\ \mathbf{u}_i \mid \boldsymbol{\pi})$ 具有已知函数形式，并且如果关于 $\mathbf{z}_i$ 和 $\mathbf{u}_i$，它是可加性分开的，这就能够写成：

$$\mathbf{y}_i=\mathbf{g}(\mathbf{z}_i \mid \boldsymbol{\pi})+\mathbf{u}_i=\mathrm{E}\left[\mathbf{y}_i \mid \mathbf{z}_i\right]+\mathbf{u}_i \tag{2.3}$$

那么 $\mathbf{y}$ 对 $\mathbf{z}$ 的回归就是给定 $\mathbf{z}$ 时关于 $\mathbf{y}$ 的一个自然而然的预测函数。在这个意义上，简化式方程有助于在实施给定$(\mathbf{z}_i,\ \mathbf{u}_i)$时对 $\mathbf{y}_i$ 进行条件预测。对指定式(2.2)右边变量的值来说，为了生成左边变量的预测，需要 $\boldsymbol{\pi}$ 的估计值，这在计算上是比

较简单的。

式(2.3)的一个重要推广是变换模型(**transformation model**),对于纯量 y 而言,它采用的形式为:

$$\Lambda(y)=\mathbf{z}'\boldsymbol{\pi}+\mathbf{u} \tag{2.4}$$

其中,$\Lambda(y)$表示变换函数[例如,$\Lambda(y)=\ln(y)$或 $\Lambda(y)=y^{1/2}$]。在一些情况下,变换函数可以依赖于未知参数。变换模型有别于回归,但它还是能够用于实施对$\mathrm{E}[y|\mathbf{z}]$的估计。一个重要的事例就是将在第 17 章分析的加速失败时间模型。

在对结构模型的设定中,最重要的且潜在地引起争论的步骤之一是性质 3,其中把变量分成因果与效应的先验排序就被认为是指定的。本质上,这使我们要区分两种不同的变量,一种变量的变化由设计的模型来解释,而另外一种变量的变化是由外部决定的,因此,它的变化不在我们的研究范围之内。在微观经济计量学中,前者的事例包括受教育年数与工作小时数;后者的事例包括性别、民族、年龄以及类似的人口数量。前者记为 $\mathbf{y}$,称为内生变量(**endogenous variable**),而后者记为 $\mathbf{z}$,称为外生变量(**exogenous variable**)。

变量的外生性是一个重要的简化,因为它在本质上可对下述决策判断对错,即把那些变量处理成辅助的,而不是要建立的那些变量,原因是关系函数不对研究的变量产生影响。这种重要的概念需要更正式的定义,现在我们就给出定义。

2.3 外生性

我们以考察一般有限维参数情况的表达式来开始,由于参数 $\boldsymbol{\theta}$ 分割成 $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$,$\mathbf{W}$ 的联合分布可被因式分解为给定 $\mathbf{Z}$ 时 $\mathbf{Y}$ 的条件密度以及给定

$$f_J(\mathbf{W}|\boldsymbol{\theta})=f_C(\mathbf{Y}|\mathbf{Z}, \boldsymbol{\theta})\times f_M(\mathbf{Z}|\boldsymbol{\theta}) \tag{2.5}$$

时 $\mathbf{Z}$ 的边缘密度。如果:

$$f_J(\mathbf{W}|\boldsymbol{\theta})=f_C(\mathbf{Y}|\mathbf{Z}, \boldsymbol{\theta}_1)\times f_M(\mathbf{Z}|\boldsymbol{\theta}_2)$$

那么会产生一种特殊情况,其中,$\boldsymbol{\theta}_1$ 与 $\boldsymbol{\theta}_2$ 在函数形式上是独立的。于是,我们就说,$\mathbf{Z}$ 关于 $\boldsymbol{\theta}_1$ 是外生的;这意味着,对 $\boldsymbol{\theta}_1$ 进行推断并不需要 $f_M(\mathbf{Z}|\boldsymbol{\theta}_2)$的知识,因此,我们能有效地把 $\mathbf{Y}$ 的分布以 $\mathbf{Z}$ 为条件。

一些模型总是可以重新参数化的。因此,接下来考虑用参数 φ对模型重新参数化,φ 作为对 $\boldsymbol{\theta}$ 的一一变换,比如说,$\varphi=h(\boldsymbol{\theta})$,其中,$\varphi$被分割成$(\varphi_1, \varphi_2)$。例如,如果 φ_1 对于政策干预种类而言是结构不变的,那么这种重新参数化或许是关注的内容。假定 φ_1 是关注的参数。在这种情况下,人们对 $\mathbf{Z}$ 关于 φ_1 的外生性感兴趣。于是,外生性的条件是:

$$f_J(\mathbf{W}|\varphi)=f_C(\mathbf{Y}|\mathbf{Z}, \varphi_1)\times f_M(\mathbf{Z}|\varphi_2) \tag{2.6}$$

其中,φ_1 与 φ_2 是独立的。

最后,考察关注于参数 λ 的情况,λ 是 φ 的函数,比如说 $h(\varphi)$。于是,对于 $\mathbf{Z}$

关于 λ 的外生性来说，我们需要两个条件：(i) λ 只依赖于 φ_1，也就是说，$\lambda=h(\varphi_1)$，因而，仅有条件分布是关注的内容；(ii) φ_1 与 φ_2 均是“自由变动的”(variation free)，这意味着联合分布的参数不受限于交叉约束，也就是说，$(\varphi_1, \varphi_2)\in\boldsymbol{\Phi}_1\times\boldsymbol{\Phi}_2=\{\varphi_1\in\boldsymbol{\Phi}_1,\ \varphi_2\in\boldsymbol{\Phi}_2\}$。

在外生性概念的发展中，式(2.5)与式(2.6)中的因式分解起着重要作用。本书特别关注与外生性相关的下述三个概念。

定义 2.1(弱外生性)：对于 λ 而言，$\mathbf{Z}$ 是弱外生的(**weakly exogenous**)，如果(i)与(ii)均成立。

如果边际模型参数对推断 λ 而言都是没有信息价值的，那么推断 λ 就只能根据条件分布 $f(\mathbf{Y}|\mathbf{Z},\ \varphi_1)$ 来继续进行。其运算意义为，如果人们关注于推断 λ 或 φ_1，那么弱外生变量可以取成给定的。这样做并不意味着，不存在 $\mathbf{Z}$ 的统计模型；它意味着模型的参数在推断 φ_1 时没有起作用，从而它们是无关的。

2.3.1 *条件独立性*

就起源而论，格兰杰因果性(Granger causality)概念可在时间序列背景下就预测内容来定义。更一般地讲，它能够被解释成条件独立性(**conditional independent**)的形式[霍兰德(Holland, 1986，第 957 页)]。

把 $\mathbf{z}$ 分割成两个子集 $\mathbf{z}_1$ 与 $\mathbf{z}_2$。

设

$$\mathbf{W}=[\mathbf{y}_1,\ \mathbf{z}_1,\ \mathbf{z}_2] \tag{2.7}$$

表示关注变量的矩阵。于是，给定 $\mathbf{z}_2$ 时，如果：

$$f(\mathbf{y}|\mathbf{z}_1,\ \mathbf{z}_2)=f(\mathbf{y}|\mathbf{z}_2) \tag{2.8}$$

则 $\mathbf{z}_1$ 与 $\mathbf{y}$ 是条件独立的。这一概念比均值独立性(**mean independent**)假设要强一些，它蕴含着：

$$\mathrm{E}(\mathbf{y}|\mathbf{z}_1,\ \mathbf{z}_2)=\mathrm{E}(\mathbf{y}|\mathbf{z}_2) \tag{2.9}$$

于是，一旦以 $\mathbf{z}_2$ 为条件，$\mathbf{z}_1$ 不会有关于 $\mathbf{y}$ 的预测值。在预测条件下，这意味着 $\mathbf{z}_1$ 不是 $\mathbf{y}$ 的格兰杰原因。

在时间序列背景下，$\mathbf{z}_1$ 与 $\mathbf{z}_2$ 将是 $\mathbf{y}$ 子集的互不相交滞后值。

定义 2.2 (强外生性)　如果 $\mathbf{z}_1$ 关于 φ 是弱外生性的且不是 $\mathbf{y}$ 的格兰杰原因，则 $\mathbf{z}_1$ 关于 φ 是强外生的(**strongly exogenous**)，所以式(2.8)成立。

2.3.2 *外生性变量*

外生性是一个强的假设。相对于关注参数来说，它是随机变量的性质。因此，在一个结构模型中，可以把变量有效地处理成外生的，而在另外一个模型中则不能；重要的问题是，一些参数作为推断的主题。对这个性质的任意加强将会得到某种人们满意的结果，这将在 2.4 节加以讨论。

外生性假设可以由先前的理论来判断正确，它作为维持模型假设的一部分。在一些情况下，它被证明是一种有效的近似，在这种情况下，它受限于检验，如同 8.4.3 节所讨论的。在横截面分析中，它可以被证明是一个自然实验或一个拟实

验的结果，其中，变量的值是由外部干预来确定的；例如，政府或管理机构可以确定税率的设置或者政策参数。特别关注的内容是下述情况，干预结果导致了重要政策变量值的变动。这种自然实验等同于某些变量的外生化。正如我们将在第 3 章看到的，这将创立在没有其他复杂因素的情况下去研究变量影响的准实验机会(quasi-experimental opportunity)。

2.4 线性联立方程模型

在式(2.1)中设定的一般结构模型的一个重要特殊情况，是由考尔斯委员会经济计量学家发展起来的线性联立方程模型。在许多文献中[譬如萨根(Sargan, 1988)]，可以找到对这种模型的综合处理。这里的研究是概括性的、选择性的；参见 6.9.6 节。其目的是对几种重点思想和概念进行讨论，而这些思想和概念具有更普遍的关联。尽管这样的分析局限于线性模型，但是一些见解可以被日常应用到非线性模型上。

2.4.1 *SEM 设置*

线性联立方程模型(SEM)设置如下：

$$\begin{aligned} y_{1i}\beta_{11}+\cdots+y_{Gi}\beta_{1G}+z_{1i}\gamma_{11}+\cdots+z_{Ki}\gamma_{1K}&=u_{1i}\\ \vdots \qquad\qquad\qquad \vdots \qquad &= \vdots \\ y_{1i}\beta_{G1}+\cdots+y_{Gi}\beta_{GG}+z_{1i}\gamma_{G1}+\cdots+z_{Ki}\gamma_{GK}&=u_{Gi} \end{aligned}$$

其中，i 表示观测值下标。

在内生变量 $\mathbf{y}_i'=(y_{1i},\ \cdots,\ y_{Gi})$与外生变量 $\mathbf{z}_i'=(z_{1i},\ \cdots,\ z_{Gi})$之间，做出一种清晰的先验差别或者预先安排。由定义，外生变量与纯随机扰动$(u_{1i},\cdots,\ u_{Gi})$是不相关的。在其无约束形式中，每一个变量都可以进入每一个方程。

在矩阵记号下，G 个方程 SEM 的第 i 个方程可写成：

$$\mathbf{y}_i'\mathbf{B}+\mathbf{z}_i'\mathbf{\Gamma}=\mathbf{u}_i' \tag{2.10}$$

其中，$\mathbf{y}_i$、$\mathbf{B}$、$\mathbf{z}_i$、$\mathbf{\Gamma}$ 以及 $\mathbf{u}_i$ 的维数分别为$G\times1$、$G\times G$、$K\times1$、$K\times G$ 以及$G\times1$。对于$(\mathbf{B},\ \mathbf{\Gamma})$与$(\mathbf{z}_i,\ \mathbf{u}_i)$的设定值来说，原则上能够求解出关于 $\mathbf{y}_i$ 的G 个线性联立方程。

SEM 的标准假设如下：

1. $\mathbf{B}$ 是非奇异的，且具有秩 G。
2. $\text{rank}[\mathbf{Z}]=K$。$N\times K$ 阶矩阵 $\mathbf{Z}$ 是由 $\mathbf{z}_i'$，$i=1,\ 2,\cdots,\ N$ 叠放形成的。
3. $\text{plim}\ N^{-1}\mathbf{Z}'\mathbf{Z}=\mathbf{\Sigma}_{zz}$是对称的 $K\times K$ 阶正定矩阵。
4. $\mathbf{u}_i\sim\mathcal{N}[\mathbf{0},\ \mathbf{\Sigma}]$，也就是说，$\text{E}[\mathbf{u}_i]=\mathbf{0}$ 且 $\text{E}[\mathbf{u}_i\mathbf{u}_i']=\mathbf{\Sigma}=[\sigma_{ij}]$，其中，$\mathbf{\Sigma}$ 表示对称的 $G\times G$ 阶正定矩阵。
5. 每一个方程的误差项是序列独立的。

在这种模型中，其结构(或结构参数)是由$(\mathbf{B},\ \mathbf{\Gamma},\ \mathbf{\Sigma})$构成的。一旦写成：

$$\mathbf{Y}=\begin{bmatrix}\mathbf{y}_1'\\ \vdots\\ \mathbf{y}_N'\end{bmatrix},\quad \mathbf{Z}=\begin{bmatrix}\mathbf{z}_1'\\ \vdots\\ \mathbf{z}_N'\end{bmatrix},\quad \mathbf{U}=\begin{bmatrix}\mathbf{u}_1'\\ \vdots\\ \mathbf{u}_N'\end{bmatrix}$$

这可以允许我们以更紧凑的形式把结构模型写成：

$$\mathbf{YB}+\mathbf{Z\Gamma}=\mathbf{U} \tag{2.11}$$

其中，矩阵 $\mathbf{Y}$、$\mathbf{B}$、$\mathbf{Z}$、$\mathbf{\Gamma}$ 与 $\mathbf{U}$ 的阶数分别为 $N\times G$、$G\times G$、$N\times K$、$K\times G$ 以及 $N\times G$。一旦用所有外生变量求解出所有内生变量，我们就得出 SEM 的简化式(reduced form of the SEM)：

$$\mathbf{Y}+\mathbf{Z\Gamma B}^{-1}=\mathbf{UB}^{-1}$$
$$\mathbf{Y}=\mathbf{Z\Pi}+\mathbf{V} \tag{2.12}$$

其中，$\mathbf{\Pi}=-\mathbf{\Gamma B}^{-1}$，而 $\mathbf{V}=\mathbf{UB}^{-1}$。给定假设 4，则 $\mathbf{v}_i\sim\mathcal{N}[\mathbf{0},\ \mathbf{B}^{-1\prime}\mathbf{\Sigma B}^{-1}]$。

在 SEM 框架下，有几种原因促使结构模型成为首要的形式。第一，方程本身可以拥有对诸如需求或供给关系、生产函数等经济关系的解释，而且方程受限于经济理论的约束。因此，$\mathbf{B}$ 与 $\mathbf{\Gamma}$ 都是描述经济行为的参数。因而，先验的理论能够产生关于个体系数的符号及大小的预期。与之相比，无约束简化式参数潜在地作为结构参数的复杂函数，而且照此计算它们的后估计很困难。如果经济计量建模的目标是预测，而不是对带有行为解释的参数进行推断，那么这种考虑或许不怎么重要。

为了不失一般性，考察模型(2.11)中的第一个方程，把 y_1 作为因变量。此外，剩下的 $G-1$ 个内生变量与 $K-1$ 个外生变量的一部分可以不在这个方程之中。从式(2.12)中我们看到，内生变量 $\mathbf{Y}$ 通常随机地依赖于 $\mathbf{V}$，同样也是结构误差 $\mathbf{U}$ 的函数。因此，通常 $\text{plim}\ N^{-1}\mathbf{Y}'\mathbf{U}\neq\mathbf{0}$。一般地，对联立方程应用最小二乘法估计，会产生非一致性估计值。这是众从周知的且基本的结果，它起源于联立方程文献，经常被称为“联立方程偏倚”的问题。当最小二乘方法失效时，有大量的联立方程模型的文献处理识别与一致性估计；参见萨根(Sargan, 1988)和施密特(Schmidt, 1976)以及 6.9.6 节。

SEM 的简化式把每一个内生变量表示成所有外生变量与所有结构扰动项的线性函数。简化式扰动项是结构扰动项的线性组合。出自简化式的第 i 个观测值为：

$$\mathrm{E}[\mathbf{y}_i\,|\,\mathbf{z}_i]=\mathbf{z}_i'\mathbf{\Pi} \tag{2.13}$$
$$\mathrm{V}[\mathbf{y}_i\,|\,\mathbf{z}_i]=\mathbf{\Omega}\equiv\mathbf{B}^{-1\prime}\mathbf{\Sigma B}^{-1} \tag{2.14}$$

简化式参数 $\mathbf{\Pi}$ 可以被推导成为结构参数的函数所定义的参数。如果 $\mathbf{\Pi}$ 能够被一致估计，那么简化式能做出关于 $\mathbf{Y}$ 随 $\mathbf{Z}$ 中外生变化而变化的预测陈述。即使 $\mathbf{B}$ 与 $\mathbf{\Gamma}$ 都是未知的，这也是可能的。给定 $\mathbf{Z}$ 的外生性，简化式回归的全部集合是多变量回归模型，这可以通过最小二乘法得到一致估计。简化式为给定 $\mathbf{Z}$ 时对 $\mathbf{Y}$ 进行条件预测提供了基础。

约束简化式是由无约束简化式模型受限于约束而得到的。如果这些约束与那些应用到结构式上的一样，那么结构信息就能够被简化式重新利用。

在 SEM 框架下，未知结构参数 $\mathbf{B}$、$\mathbf{\Gamma}$ 以及 $\mathbf{\Sigma}$ 的非零元素起着核心作用，因为它们反映了模型的因果结构。内生变量之间的相互依存是由 $\mathbf{B}$ 来刻画的，而内生变

量对 Z 中的外生冲击的响应则反映在参数矩阵中。在这种设置下，关注的因果参数就是那些可测量解释变量 y_i 或 z_k 的变化对关注结果 y_l，$l \neq j$ 的直接边际影响，以及这些参数与数据的函数。$\mathbf{\Sigma}$ 元素刻画出随机扰动项的离散趋势与相关性，因此，它们测量数据生成方式的某些性质。

2.4.2 *SEM 的因果解释*

用一个简单的事例来阐明 SEM 中参数的因果解释。结构模型具有两个连续内生变量 y_1 与 y_2、一个连续外生变量 z_1、一个联系 y_1 与 y_2 的随机关系，以及一个联系模型中全部三个变量的可定义的恒等式：

$$\begin{aligned} y_1 &= \gamma_1 + \beta_1 y_2 + u_1, \quad 0<\beta_1<1 \\ y_2 &= y_1 + z_1 \end{aligned}$$

在此模型中，u_1 表示随机扰动项，它与 z_1 是独立的，具有定义良好的分布。参数 β_1 受限于同样作为该模型设定一部分的一个不等式约束。变量 z_1 是外生的，因此它的变动是由我们认为是干预的外部来源而引起的。这些干预通过恒等式对 y_2 具有直接影响，而且通过第一个式子对其产生间接影响。影响可通过该模型的简化式来测量，即：

$$\begin{aligned} y_1 &= \frac{\gamma_1}{1-\beta_1} + \frac{\beta_1}{1-\beta_1} z_1 + \frac{1}{1-\beta_1} u_1 \\ &= E[y_1 \mid z_1] + v_1 \\ y_2 &= \frac{\gamma_1}{1-\beta_1} + \frac{1}{1-\beta_1} z_1 + \frac{1}{1-\beta_1} u_1 \\ &= E[y_2 \mid z_1] + v_1 \end{aligned}$$

其中，$v_1 = u_1/(1-\beta_1)$。简化式系数 $\beta_1/(1-\beta_1)$ 与 $1/(1-\beta_1)$ 具有因果解释。任何外部引起的 z_1 上的变动都将引发 y_1 与 y_2 变动这些数量。注意到，在这个模型中，y_1 与 y_2 也对 u_1 有响应。为了不混淆对这两个变动来源的影响，我们要求 z_1 与 u_1 是独立的。

同样注意到：

$$\begin{aligned} \frac{\partial y_1}{\partial y_2} &= \beta_1 = \frac{\beta_1}{1-\beta_1} \div \frac{1}{1-\beta_1} \\ &= \frac{\partial y_1}{\partial z_1} \div \frac{\partial y_2}{\partial z_1} \end{aligned}$$

在什么意义下，β_1 能测度 y_2 对 y_1 的因果效应呢？为了理解可能的困难，观察发现，y_1 与 y_2 是相互依存的或是联合确定的，因此，在什么意义下 y_2"引起"y_1 并不清楚。虽然 z_1（以及 u_1）在简化式意义下是变动的最终原因，但 y_2 是 y_1 的近似原因或者中间原因。也就是说，第一个结构方程提供了 y_2 对 y_1 影响的简要描述，而简化式则为了计算它们而给出了在考虑两个内生变量之间的所有交互作用之后的（均衡）影响。在 SEM 框架下，甚至把内生变量看成是因果变量，而且把它们的系数作为因果系数。这一方法能够产生困惑，即在以变动的独立来源作为因果变

量的实验背景下，把谁看成因果性。如果 y_2 是独立的且是外生来源，那么 SEM 方法有意义，在此模型中变动就是 z_1。因此，边际响应系数 β_1 是 y_1 与 y_2 如何响应 z_1 变化的函数，因为前面的方程是清楚的。

当然，这个模型只是一种特殊情况。更一般地讲，我们可以询问在什么条件下，SEM 参数将具有有意义的因果解释。在 2.5 节讨论识别概念时，我们将回到这个问题上。

2.4.3 扩展到非线性和潜变量模型

如果联立模型仅仅关于参数是非线性的(**nonlinear in parameters**)，那么结构模型可以写成：

$$\mathbf{YB}(\boldsymbol{\theta})+\mathbf{Z\Gamma}(\boldsymbol{\theta})=\mathbf{U} \tag{2.15}$$

其中，$\mathbf{B}(\boldsymbol{\theta})$ 与 $\boldsymbol{\Gamma}(\boldsymbol{\theta})$ 都表示矩阵，它们的元素均是结构参数 $\boldsymbol{\theta}$ 的函数。如上所述，能够推导出显性的简化式。

然而，如果非线性(**nonlinear**)是关于变量(**in variables**)的，那么获得显性(解析的)简化式是不可能的，尽管给定($\mathbf{z}$, $\mathbf{u}$)时通常能获得因变量的线性化近似或者数值解。

许多微观经济计量模型都涉及潜变量[1](**latent variables**)或者不可观测变量(**unobserved variables**)，以及可观测内生变量(**observed endogenous variables**)。例如，搜索和拍卖理论模型均使用保守工资或保留价格的概念，选择模型引起了间接效用等。在这类模型情况下，结构模型(2.1)可以由

$$\mathbf{g}(\mathbf{y}_i^*,\ \mathbf{z}_i,\ \mathbf{u}_i\,|\,\boldsymbol{\theta})=\mathbf{0} \tag{2.16}$$

代替，其中，潜变量 $\mathbf{y}_i^*$ 代替可观测变量 $\mathbf{y}_i$。用($\mathbf{z}_i$, $\mathbf{u}_i$)来求解关于 $\mathbf{y}_i^*$ 的对应简化式，得到：

$$\mathbf{y}_i^*=\mathbf{f}(\mathbf{z}_i,\ \mathbf{u}_i\,|\,\boldsymbol{\pi}) \tag{2.17}$$

由于 $\mathbf{y}_i^*$ 并不是完全可观测的，所以这种简化式具有有限的作用。然而，如果我们具有函数 $\mathbf{y}_i=\mathbf{h}(\mathbf{y}_i^*)$，它与可观测的 $\mathbf{y}_i$ 的潜在部分相关联，那么用可观测的形式表示简化式就是：

$$\mathbf{y}_i=\mathbf{h}(\mathbf{f}(\mathbf{z}_i,\ \mathbf{u}_i\,|\,\boldsymbol{\pi})) \tag{2.18}$$

对于更进一步内容，可参见 16.8.2 节。

当结构模型包括关于变量的非线性或者当其涉及潜变量时，很难得到这种简化式函数形式的显性推导。在一些情况下，实践者使用近似。通过运用数学上或计算上的便利，一种特定的函数形式可以把内生变量与所有的外生变量联系起来，而这一结果称为“简化式类型关系”。

---

[1] 又称为潜在变量。——译者注

2.4.4 *对结构关系的解释*

马歇克(Marschak, 1953, 第 26 页)在一篇有影响的论文中,给出结构的下述定义:

> 结构被定义为一系列不变化的条件,尽管有观测发生,但将会变动。如果设定的结构变化如同人们期望的或者打算的,对于政策制定者来说,关注变量的预测就需要过去结构的知识……在经济学中,构成结构的条件是:(1) 刻画人类行为与制度以及技术规律的一系列关系,通常会涉及不可观测的随机扰动项和不可观测的测量误差;(2) 这些随机量的联合概率分布。

马歇克认为,结构是定量评估或检验经济理论的基础,同时最佳政策的选择需要结构的知识。

在 SEM 文献中,结构模型意指"自治的"(不是推导的)关系。存在着另外一些与结构密切相关的概念。一种这样的概念涉及"深参数",它意指那些对于干预而言不变的技术和偏好参数。

在最近几年,对结构术语的使用出现了一种可供选择的形式,即经济计量模型建立在由理性行为人引起的动态随机最优化的假设基础之上。这种方法中,对任何结构进行估计的问题出发点都是去定义行为人最优化行为的一阶必要条件。例如,约束最大化效用的标准问题中,行为关系是确定性一阶边际效用条件。如果有关的函数形式都能得到显性表述,并引进最优化的随机误差,那么一阶条件定义一种行为模型,其参数可以刻画效用函数——所谓的深参数或政策不变参数。一些事例将在 6.2.7 节和 16.8.1 节给出。

这种高度结构化方法(**highly structured approach**)的两个特点应该提及。第一,该方法以严谨方式依赖于先前的经济理论。经济理论不能用于直接生成人们以在一定程度上任意设置的函数形式来使用的一系列相关变量。相反,基本经济理论在设定、估计以及推导中起着重要作用(但不是排他的)。第二,所得到的模型的估计、识别以及设定可以相当复杂,因为行为人的最优化问题可能非常复杂,特别是当不确定性下的动态最优化被假定,并且离散性与非连续性都得以表现出来时;参见拉斯特(Rust, 1994)。

2.5 识别概念

SEM 方法的目的是要一致地估计出($\mathbf{B}$, $\mathbf{\Gamma}$, $\mathbf{\Sigma}$),并且进行统计推断。一致估计的一个重要预先条件是,模型应该是可识别的。我们将在参数模型背景下简要地讨论以下两个相关的重要概念:观测等价性(**observational equivalence**)与可识别性(**identifiability**)。

识别与给定充分观测值时参数的确定有关。在这个意义上,它是一个渐近概念。统计上的不确定性必定要影响到建立在有限观测值基础上的任何判断。由基于假定的考虑,有限多个观察值是可以利用的。可以考虑在其点值意义上,或者在

确定那些参数成为元素的集合意义上，决定一个关注参数是否是逻辑上可行的。因此，识别是一种基础性考虑，而且在逻辑上应优于统计估计，并与统计估计相互分开，这一点也是本节要强调的。然而，集合识别（**set identification**）或者界限识别（**bounds identification**）是一种重要的方法，这种方法在本书的个别地方将会用到[例如，第 25 章和第 27 章；参见曼斯基（Manski，1995）]。经济计量文献中，有关识别的大量内容都关注于点识别。

定义 2.3（观测等价性） 一个模型所定义的联合概率分布函数 $\Pr[\mathbf{x}|\boldsymbol{\theta}]$，$\mathbf{x}\in\mathbf{W}$，$\boldsymbol{\theta}\in\boldsymbol{\Theta}$ 的两个结构在观测上是等价的，如果 $\Pr[\mathbf{x}|\boldsymbol{\theta}^1]=\Pr[\mathbf{x}|\boldsymbol{\theta}^2]$，$\forall\mathbf{x}\in\mathbf{W}$。

稍微正式地讲，如果给定数据，两个结构模型意味着相同的联合概率分布，那么，这两个结构在观测上就是等价的。在观测上等价的多重结构的存在意味着识别失效。

定义 2.4（识别性） 一个结构 $\boldsymbol{\theta}^0$ 是可识别的，如果在 $\boldsymbol{\Theta}$ 上不存在观测上等价的其他结构。

在线性回归 $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{u}$ 中，当回归元之间存在完全多重共线性时，就是一个简单的非识别事例。于是，我们能识别线性组合 $\mathbf{C}\boldsymbol{\beta}$，其中，$\text{rank}[\mathbf{C}]<\text{rank}[\boldsymbol{\beta}]$，但是，我们却不能识别 $\boldsymbol{\beta}$ 本身。

这种定义涉及结构的唯一性。在我们已给定的 SEM 背景下，这一定义意味着，识别要求存在唯一的三元组($\mathbf{B}$，$\boldsymbol{\Gamma}$，$\boldsymbol{\Sigma}$)与已观测的数据一致性。在 SEM 中，如同在其他情况中一样，识别涉及能获得给定数据的样本矩时结构参数的唯一估计值。例如，就简化式(2.12)而言，在所述假设下，最小二乘法估计量提供 $\boldsymbol{\Pi}$ 的唯一估计值，也就是说，$\hat{\boldsymbol{\Pi}}=[\mathbf{Z}'\mathbf{Z}]^{-1}\mathbf{Z}'\mathbf{Y}$，并且 $\mathbf{B}$、$\boldsymbol{\Gamma}$ 的识别要求，给定模型上的一个先验约束时，出自方程 $\boldsymbol{\Pi}+\boldsymbol{\Gamma}\mathbf{B}^{-1}=\mathbf{0}$ 的 $\boldsymbol{\Gamma}$ 与 $\mathbf{B}$ 的未知元素存在唯一解。唯一解则蕴含着该模型的恰好识别。

一个完全模型被称为可识别的，如果所有模型参数都是可识别的。对于某些模型来说，仅有参数的一个子集是可识别的，这一点可能的。在一些情况下，或许重要的是识别参数的某一函数，而不必识别所有参数。参数函数的识别意味着，那个函数能够唯一地被 $F(\mathbf{W}|\boldsymbol{\Theta})$ 重新利用。

人们如何确保可以"剔除"可供选择模型设定的结构呢？在 SEM 中，对这一问题的解答依赖于通过($\mathbf{B}$，$\boldsymbol{\Gamma}$，$\boldsymbol{\Sigma}$)上的先前约束而增大的样本信息。这种先前约束必须把充分的附加信息引入模型中，以便剔除其他观测上等价结构的存在性。

下述的讨论可以证明对先前约束的需要。注意到，给定 2.4.1 节的假设，由($\boldsymbol{\Pi}$，$\boldsymbol{\Omega}$)定义的简化式总是唯一的。首先，假定($\mathbf{B}$，$\boldsymbol{\Gamma}$，$\boldsymbol{\Sigma}$)上没有约束。其次，假定存在两个观测上等价的结构($\mathbf{B}_1$，$\boldsymbol{\Gamma}_1$，$\boldsymbol{\Sigma}_1$)与($\mathbf{B}_2$，$\boldsymbol{\Gamma}_2$，$\boldsymbol{\Sigma}_2$)。于是有：

$$\boldsymbol{\Pi}=-\boldsymbol{\Gamma}_1\mathbf{B}_1^{-1}=-\boldsymbol{\Gamma}_2\mathbf{B}_2^{-1} \tag{2.19}$$

令 $\mathbf{H}$ 表示一个 $G\times G$ 阶非奇异矩阵。从而，$\boldsymbol{\Gamma}_1\mathbf{B}_1^{-1}=\boldsymbol{\Gamma}_1\mathbf{H}\mathbf{H}^{-1}\mathbf{B}_1^{-1}=\boldsymbol{\Gamma}_2\mathbf{B}_2^{-1}$，这意味着 $\boldsymbol{\Gamma}_2=\boldsymbol{\Gamma}_1\mathbf{H}$，$\mathbf{B}_2=\mathbf{B}_1\mathbf{H}$。因此，第二个结构是对第一个结构的线性变换。

对这个问题的 SEM 求解是要引入($\mathbf{B}$，$\boldsymbol{\Gamma}$，$\boldsymbol{\Sigma}$)上的一些约束，使得我们能够剔除导致观测上等价的结构的那种线性变换的存在性。换句话说，($\mathbf{B}$，$\boldsymbol{\Gamma}$，$\boldsymbol{\Sigma}$)上的约

束必须是不存在矩阵 **H**,而 **H** 会产生具有相同简化式的另外一种结构;给定($\mathbf{\Pi}$, $\mathbf{\Omega}$),将存在求解方程 $\mathbf{\Pi}=\mathbf{\Gamma}\mathbf{B}^{-1}$ 和 $\mathbf{\Omega}\equiv(\mathbf{B}^{-1})'\mathbf{\Sigma}\mathbf{B}^{-1}$ 的唯一解。

实际上,可以施加各种各样的约束,包括:(1) 正规化,例如,令 **B** 的对角元素为 1;(2) 0(排除在外)与线性齐次的以及非齐次的约束;(3) 协方差与不等式约束。在线性与非线性模型中,关于识别的必要且充分条件的详细内容,能够在许多文献中找到,包括萨根(Sargan, 1988)。

对识别约束进行有意义的利用,要求所施加的先前约束应该是后验有效的。这一思想在考虑识别问题的几章中会进一步地展开(例如,参见 6.9 节)。

排除性约束(**exclusion restrictions**)本质上表明,模型包括了对某些内生变量具有 0 影响的一些变量。也就是说,某个因果方向先前被剔除了。这使得识别因果的其他方向成为可能。

例如,在前面给定的一个简单的两个变量事例中,z_1 没有进入 y_1 方程中,使得识别 y_2 对 y_1 的直接影响成为可能。尽管排除性约束应用起来最简单,但在参数模型中识别还是可以通过不等式约束与协方差约束来得以保证。

如果 $\mathbf{\Sigma}$ 上不存在约束,且 **B** 的对角元素正规化为 1,那么关于识别的必要条件(**necessary condition**)是阶条件(**order condition**),这表示被排除的外生变量的个数必须至少等于所包括的内生变量的个数。在许多文献中都给定,充分条件(**sufficient condition**)是秩条件(**rank condition**),这保证第 j 个方程参数 $\mathbf{\Pi}\mathbf{\Gamma}_j=-\mathbf{B}_j$ 产生给定 $\mathbf{\Pi}$ 时($\mathbf{\Gamma}_j$, $\mathbf{B}_j$)的唯一解。

已知识别,恰好识别[**just (exact) identification**]术语是指阶条件得以准确满足的情况;过度识别(**overidentification**)是指,方程组上的约束个数超过恰好识别所需要的个数。

萨根(Sargan, 1988)曾经讨论非线性 SEM 中的识别,他还给出与之相关的早期参考文献。

2.6 单方程模型

为了不失一般性,考察约束于正规化 $\beta_{11}=1$ 的线性 SEM 的第一个方程。设 $y=y_1$,令 $\mathbf{y}_1$ 表示 $\mathbf{y}$ 的内生成分而不是 y_1,并且令 $\mathbf{z}_1$ 表示 $\mathbf{z}$ 的外生成分,满足:

$$y=\mathbf{y}_1'\boldsymbol{\alpha}+\mathbf{z}_1'\boldsymbol{\gamma}+u \tag{2.20}$$

许多形式都遗漏了涉及从方程组到单方程的正式步骤,并且从回归方程

$$y=\mathbf{x}'\boldsymbol{\beta}+u$$

开始,其中,$\mathbf{x}$ 的某些成分是内生的(以显性方式 $\mathbf{y}_1$),而其他一些成分是外生的(以显性方式 $\mathbf{z}_1$)。于是,关注的内容是去估计依赖于重要回归元变动的影响,该重要回归元可以是内生的也可以是外生的,这一点依赖于假设。工具变量或者两阶段最小二乘法估计是最显然的估计策略(参见 4.8 节、6.4 节以及 6.5 节)。

在 SEM 方法中,至少去设定模型中某些剩余的方程,这是很自然的,即使剩余方程并不是研究的关注点。假定 $\mathbf{y}_1$ 具有维数 1。于是,第一种可能性是去设定关

于 y_1 的结构方程，以及在关于 y_1 的这个结构方程中会出现的其他内生变量的结构方程。第二种可能性是去设定关于 y_1 的简化式方程。这将证明，影响 y_1 的外生变量并不直接影响 y。一个优点是，在这样的设置下，工具变量自然会出现。然而，在最近的实证研究工作中，在单方程设置中利用工具变量，甚至要避免写出关于右边内生变量的简化式正式步骤。

2.7 潜在结果模型

当关注于公共政策的影响和/或关于某种特定结果的私人决策变量时，经济计量模型中因果推断的动机就特别强。一些特定的事例包括转移支付对劳动力供给的影响、班级大小对学生学习的影响，以及健康保险对保健的影响。在许多情况下，因果变量本身就反映个体决策，因此，因果变量潜在地是内生的。正如通常情况一样，当经济计量估计与推断建立在观测数据（**observational data**）基础之上，就对因果参数的识别与推断提出了许多挑战。如果因果问题是利用出自受控的正常统计设计的社会实验（**social experiment**）的数据，那么这些挑战潜在地缺乏严谨性。尽管这类实验可以实施（参见 3.3 节的例子及详细内容），一般来说，实验的组织与执行是昂贵的。因此，更引人注目的是，利用由自然实验（**natural experiment**）所生成的数据或者在拟实验设置下实施因果建模。3.4 节讨论这些数据结构的优缺点；如果没有现存目的，人们就应该把自然实验或拟实验看成是一种设置，即某个因果变量外生的变动，且与其他解释变量独立，这样会使得识别因果参数相对容易一些。

因果性建模的主要障碍出于因果推断的基本问题（fundamental problem of causal inference）[霍兰（Holland, 1986）]。设 X 表示已假定的原因，而 Y 表示结果。通过对 X 值的操作，我们就能改变 Y 的值。假定 X 的值从 x_1 变动到 x_2。于是，通过比较 Y 的两个值来测量该变动对 Y 的因果影响：y_2 是由该变动引起的，而 y_1 是 x 没有发生变动时所具有的结果。然而，如果 X 为变动的，那么 Y 的值在缺乏 X 变动下是不会被观测到的。因此，在缺乏 X 变动的情况下，如果没有关于 Y 具有什么值的假设，对于因果影响就没有什么可讲的。后者可称为反事实[1]（**counterfactual**），意味着假设不可观测的值。简略地讲，所有因果推断都涉及一个事实与一个反事实结果的比较。在传统经济计量模型（比如，SEM）中，并不需要以显性方式表述反事实。

在微观经济计量文献中，一个相对比较新的领域是项目评估（**problem evaluation**）或者处置评估（**treatment evaluation**），它提供了估计因果参数的统计框架。在统计文献中，这种框架还统称为鲁宾因果模型（**Rubin causal model, RCM**），这样做是承认鲁宾早期的重要贡献，而鲁宾也是从该方法创始人 R. A. 费希尔（R. A. Fisher）那里引用的。尽管依照最近的习惯，我们把这称为鲁宾因果模型，但是斯普拉瓦—内曼（Splawa-Neyman）在 1923 年以波兰文发表的一篇文章中同样提出

〔1〕 又称为反事实框架。——译者注

了一种类似的统计模型；参见内曼（Neyman, 1990）。在经济计量学中，涉及反事实的模型则是沿着罗伊（Roy, 1951）的原创研究工作而独立发展起来的。本节剩下的内容将分析 RCM 的显著特征。

建立在反事实基础之上的因果参数提供了因果性的统计意义与操作意义，这在许多方面有别于传统的考尔斯基本定义。第一，在理想设置下，此框架导致一些经济计量方法相当简单。第二，一般地讲，此框架只关注少数几个（fewer）被认为与所要检验的政策问题最为相关的因果参数。这与同时关注于所有结构参数的传统经济计量方法形成对比。第三，该方法提供了对由标准结构方法所估计的因果参数性质的其他见解。

2.7.1 鲁宾因果模型

"处理"这一术语与"原因"可以交换使用。在医学新药评估的研究中，涉及那些接受治疗与那些没有接受治疗的组，已治疗组的药物反应与那些未治疗组的情况相比较。对因果影响的测量是已治疗组结果与未治疗组情况的平均差。在经济学中，"处理"这一术语使用得非常广泛。它涵盖了对某种结果有影响的所有变量，而这里的某种结果就是研究的目标。处理结果时的一些事例包括受教育与工资、班级大小与学业成绩、职业培训与收入。注意，处理不要求是外生的，并且在许多情况下，它是内生的（选择）变量。

在潜在结果模型（**potential outcome model, POM**）框架下，假定对象总体的每一个元素潜在地面临处理，三元组 (y_{1i}, y_{0i}, D_i)，$i=1,\cdots,N$ 构成处理评估的基础。当接受处理或未接受处理时，类别变量 D 分别取值 1 与 0；y_{1i}测度个体 i 接受处理的响应，而 y_{0i}测度未接受处理时的情况。也就是说：

$$y_i=\begin{cases} y_{1i}, & 如果\ D_i=1 \\ y_{0i}, & 如果\ D_i=0 \end{cases} \tag{2.21}$$

对于个体 i 来说，由于接受处理与未接受处理是互斥的表述，所以对任何给定的 i，两个测量仅有一个是可以利用的，没有利用的测量是反事实。原因 D 对个体 i 的结果效应是由 $(y_{1i}-y_{0i})$ 来计算的。$D_i=1$ 相对于 $D_i=0$ 的平均因果效应是由平均处理效应（**average treatment effect, ATE**）来计算的：

$$\text{ATE}=\text{E}[y|D=1]-\text{E}[y|D=0] \tag{2.22}$$

其中，期望与对象总体的概率分布有关系。与强调边际效应的传统结构不同，POM 框架则强调 ATE 以及与之相关的参数。

对 ATE 类型参数进行估计的实验方法，涉及通过对作为控制的未处理情况集合的结果进行比较而引起的随机指派（**random assignment**）。这类实验设计将在第 3 章中以更详细的方式加以阐述解释。随机指派蕴含着面临处理的个体被随机选取，因此，处理指派并不依赖于结果，而且与被处理问题的属性是不相关的。有两种简化方法如下。如果某些相关变量不可避免地从回归中被省略掉，处理变量就被处理成外生的，而且其在线性回归中的系数将不受到省略变量的偏倚的妨碍。

在某种条件下，即在第 3 章与第 25 章将以更长篇幅讨论的条件下，在被处理组的结果与控制组之间的平均将提供 ATE 的估计。对设计良好的回报就是对于所做出的那种因果表述具有相对的简单性。当然，为了确保处理效应估计具有高度的统计精度，人们还应该控制同样影响到结果的那些属性。

由于处理的随机指派一般来讲在经济学中是不可行的，所以对 ATE 类型参数进行估计，必须建立在非随机处理指派生成的观测数据基础上。因而，对 ATE 进行一致估计将会受到几个困难的威胁，例如，这些威胁包括结果在处理之间的可能相关性、省略变量以及处理变量内生性。一些经济计量学家曾建议，缺乏随机化构成了获得令人信服的有关因果关系的统计推断的主要障碍。

如果反事实能够得到清楚的表述并得到证实，那么潜在结果可以产生因果陈述。对反事实进行明显的陈述是这种模型的重要特征，对什么应该加以比较的含义清晰可见。如同具有观测数据的情况一样，如果缺少被观测的量与反事实的量之间清晰的差别，那么对谁被处理影响了的问题仍是不清楚的。ATE 是一种对特定子总体的边际响应加权并组合的测量。特定的假设要求实施反事实。对于能够观测到的被处理单位的信息，以及未被处理单位的信息，都需要去估计 ATE。例如，如果处理不能被应用，那么必须要识别代表被处理组的未被处理组。这种步骤总是能够得以执行，这样的要求不一定正确。在选择处理准确方式上，涉及在第 3 章和第 25 章将要讨论的抽样设计问题。

POM 的第二个有用特征是，它可以识别由自然实验或拟实验所产生的因果建模的机会。当数据在这种设置下得以生成时，而且倘若某些其他条件得到满足，如果没有 SEM 框架的全部复杂性，就会产生因果建模。这一问题在第 3 章和第 25 章将会进一步得到分析。

第三，与 SEM 结构形式——所有变量除被解释变量以外都能够被标记为“原因”——不同，在 POM 中，并不是所有解释变量都能够被当作原因的。许多都是在回归分析中必须加以控制的单元属性，而属性不是原因[霍兰(Holland，1986)]。表示原因的参数必须是相对于那种实际或潜在地、直接或间接地受限于干预的变量而言的。

最后，ATE 参数的可识别性或许是比较容易研究的目标，因此，在不是完全 SEM 的可识别性的一些地方，ATE 参数可识别性是可行的[安格里斯特(Angrist，2001)]。是否如此，必须在逐一情况的基础上进行确定。然而，许多可利用的 POM 应用，典型地使用有限的而不是完全信息框架。然而，甚至在 SEM 框架内，正如前面曾讨论的，使用有限的信息框架同样是可行的。

2.8 因果建模及估计策略

在本节，我们简略地叙述经济计量学家以许多不同的途径对因果关系进行建模的方法。这些方法既可用于 SEM 框架内，又可用于 POM 框架内，但对于前者而言，这些方法一般是可识别的。

2.8.1 识别框架

完全信息结构模型

这种方法的一种变性建立在以外生变量为条件的内生变量的联合分布参数的设定基础上。一些关系不一定要从最优化的行为模型中推导出来。设置参数约束是为了确保模型参数的识别，模型参数是统计推导的目的。整个模型可利用极大似然法或基于矩的估计来得到联合估计。我们称这种方法为完全信息结构方法(**full-information structural approach**)。对于设定良好的模型来说，这是一种吸引人的方法，但是，通常其潜在局限性是，它可以包括某些设定欠佳的方程。

在统计形式上，我们可以将完全信息方法解释成为，给定外生变量、内生变量的联合概率分布而形成的对因果性进行推断的基础的一种方法。接合点是从内生变量之间或内生变量与方程扰动项之间的同期相互依存性或动态相互依存性推导出的。

有限信息结构模型

与之相比，当统计推导的中心目标是对一个或者两个重要参数进行估计，就可以使用有限信息(**limited-information**)方法。这种方法的一个特征是，尽管一个方程是推断的中心，但是可利用该方程与其他内生变量之间的联合依赖性。这就需要关于模型某些特征的并不作为推断的主要目标的显性假设。工具变量法、序贯多步法以及有限信息极大似然法都是这一方法的特例。为了实施该方法，人们一般用一个(或多个)结构方程与一些以隐性或显性方式表述出的简化方程进行。这与全部方程都为结构的完全信息方法形成了对照。有限信息方法在计算机处理方面常常比完全信息方法更容易。

在统计形式上，我们将有限信息方法解释为，把联合分布因式分解为所关注的内生变量(比如说 $\mathbf{y}_1$)的条件模型与其他内生变量(比如说 $\mathbf{y}_2$)的边际模型的乘积的一种方法，它是一些条件变量的集合，如同：

$$f(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})=g(\mathbf{y}_1|\mathbf{x},\mathbf{y}_2,\boldsymbol{\theta}_1)h(\mathbf{y}_2|\mathbf{x},\boldsymbol{\theta}_2),\quad \boldsymbol{\theta}\in\boldsymbol{\Theta} \tag{2.23}$$

如果 $\boldsymbol{\theta}_2$ 被认为是冗余参数(**nuisance parameters**)，那么模型就可以建立在对 $h(\mathbf{y}_2|\mathbf{x},\boldsymbol{\theta}_2)$最少关注的 $g(\mathbf{y}_1|\mathbf{x},\mathbf{y}_2,\boldsymbol{\theta}_1)$成分基础上。当然，这种因式分解不是唯一的，因而，有限信息方法能够有几种变化形式。

可识别的简化式

SEM 方法的第三种变化形式是以可识别简化式来进行的。这里，人们还是对结构参数感兴趣。然而，从受限于约束的简化式中去估计结构参数是方便的。在时间序列中，可识别向量自回归提供了一个事例。

2.8.2 识别策略

存在许多潜在途径，使得对重要模型参数的识别受到危害。省略变量、函数形式错误设定、解释变量中的测量误差、利用总体的非代表性数据，以及忽略解释变量的内生性，都是一些重要的事例。微观经济计量学包括了如何解决这些问题的

许多特定事例。安格里斯特和克鲁格(Angrist and Krueger，2000)曾提供对劳动经济学中普遍流行的识别策略的一个综述，强调了 POM 框架。本书其他地方对许多这类问题加以发展，但这里仍是简略的提及。

外生化

数据有时是由自然实验与拟实验生成的。这里的思想就是，政策变量对于某些子总体而言可以是外生的变动，而它对于其他子总体而言仍是相同的。例如，最低工资法在一个州内可以是变动的，而在其邻近州内仍保持不变。这样的事件自然创造了处理组与控制组(对照组)。如果自然实验近似于一个随机化处理安排，那么利用这种数据去估计结构参数，比对含有内生处理变量的较大联立方程模型进行估计更简单。还有一种可能，自然实验中的处理变量可以被看成外生的，但处理本身却不是随机指派的。

剔除冗余函数

在有大量冗余函数的情况下，识别会受到威胁。例如，在横截面回归模型中，条件均值函数 $E[y_i|\mathbf{x}_i]$ 可以包括特定个体固定效应 $\boldsymbol{\alpha}_i$，假定与回归误差是相关的。如果每一个个体都没有许多观测值(比如，面板数据)，这一效应就不能是可识别的。然而，带有短面板的数据，通过模型变换能够把固定效应剔除掉。另外一个事例是存在不随时间而变并且不可观测的外生变量，该外生变量对一些个体组而言却是共同的。在剔除固定效应的变换事例中，通过对模型取差分和差分的差分来进行。

控制混淆

当一些变量从回归中被省略掉，并且当省略因素与包含的变量相关时，混淆偏倚就产生了。例如，在把收入作为因变量与把受教育作为解释变量的回归中，个人能力可以被看成是被省略的变量，因为它通常仅仅是不完美代表。这意味着受教育变量的系数潜在地不是可识别的。一种可行的策略是要在模型中引入控制变量(**control variables**)；这一通用方法称为控制函数方法(**control function approach**)。这些变量试图去逼近省略变量的影响。例如，各种形式的学术成就得分可以作为对能力的控制。

创建综合样本

在 POM 框架下，因果参数可以是不可识别的，因为没有合适的比较或者对照组(控制组)去提供关于估计的一个基准。潜在解决方法是创建一个综合样本，它包括作为代表控制的比较组。这种样本是由配对(**matching**)创建的(在第 25 章讨论)。如果处理组能被良好配对控制所扩大，那么对因果参数的识别在与 ATE 相关的参数得到估计的意义上可以完成。

工具变量

如果因为处理变量是内生的，识别处于危及境地，那么一种标准的解决方法是使用有效的工具变量。说比做更容易一些。对工具变量的选择是敏感的。在 4.8 节、4.9 节、6.4 节、6.5 节和 25.7 节，以及书中的其他几个地方，都对该方法进行了分析。另一方面，自然实验可以提供有效的工具。

重新对样本加权

如果样本数据是总体的代表，那么以此样本为基础对总体进行推断才是有效

的。当样本数据不具有代表性时，就产生样本选择问题或有偏抽样，此时总体参数是不可识别的。要解决这种问题，需要对样本选择进行修正的方法（第16章），或者需要对样本信息重新加权的方法。

2.9 文献注释

2.1 由赫克曼和麦克法登所提供2001年诺贝尔奖演讲稿，是关于微观经济计量学发展的既具有历史信息又具有当前信息的珍贵资料。赫克曼的演讲着重于对其综合概述，并且对微观经济计量学的许多方面提出了相当多的见解。他对异质性的讨论有许多要点与本书所涵盖的几个专题相联系。

2.2 马歇克（Marschak，1953）给出关于对政策评价的结构建模的最初经典陈述。他很早就提及参数不变性的思想。

2.3 恩格尔、亨德里和理查德（Engle，Hendry，and Richard，1983）曾提出，利用可观测变量的分布来定义弱外生性与强外生性。他们把先前文献中关于外生性的一些概念联系起来。

2.4和2.5 "识别"术语是库普曼（Koopmans，1949）使用的。大部分教科书都涵盖了线性参数模型中的点识别，包括由萨根（Sargan，1988）给出的一种综合而简明的研究、戴维森和麦金农（Davidson and Mackinnon，2004），以及格林（Greene，2003）的著作。古里耶克斯和蒙福特（Gourièroux and Monfort，1989，第3章、第4章）提供一种利用费希尔和库尔贝克（Fisher and Kullback）信息测量的不同观点。在几种重要情况下，界限识别是由曼斯基发展起来的（Manski，1995）。

2.6 赫克曼（Heckman，2000）提供传统经济计量学模型中因果性的历史概览和现代解释。在POM框架下的因果概念是由霍兰（Holland，1986）仔细而深刻地进行分析，他还将其他一些定义联系起来。从历史观点来看，关于因果性的统计学家观点的一个事例是由弗里德曼（Freedman，1999）建立的。玻尔（Pearl，2000）给出将"因果关系处理成为在干预下的行为概括"思想的精彩图式阐述，并在非实验情形下推断因果关系的众多问题。

2.7 安格里斯特和克鲁格（Angrist and Krueger，1999）利用劳动经济学的事例，求解了识别陷阱。

3 微观经济数据结构

3.1 引 论

本章将概述各种类型的微观经济数据的潜在用途和局限性。微观经济计量学中最广泛使用的数据结构是调查或人口普查数据。这些数据通常称为观测数据(**observational data**),以此将它们与实验数据(**experimental data**)相区别。

本章讨论前面提及的数据结构的潜在局限性。此外,观测数据的内在局限性是以搜集数据的方式混合而成的,即通过样本框(样本生成方式)、样本设计(简单随机抽样到分层随机抽样)以及样本范围(横截面到纵向数据)的方式。因此,我们讨论与使用观测数据相关的一些抽样问题。在这一层面上,有些数据是崭新的,本章稍后对它们加以阐述。

在简单随机抽样假设下,微观经济计量学超越调查数据分析的范围。本章考察一些扩展形势。3.2 节概述多阶段样本调查的体系以及偏离随机抽样的某些普遍形式;稍后一些章节提供对它们在统计意义上的更详尽的分析。此外,考察导致数据不一定代表总体的一些复杂性。如果因果参数中缺少观测数据,就增加利用实验数据或半实验数据以及一些框架组织。3.3 节考察源自社会实验的可能性。3.4 节考察由特定的观测数据类型产生的建模机会,该特定观测数据是在半实验条件下生成的,这自然提供已处理的与未处理的对象,因此称之为自然实验。3.5 节涵盖微观数据管理的实际问题。

3.2 观测数据

微观经济观测数据的主要来源,是对住户、厂商的调查以及政府管理数据。人口普查数据也可以用于生成样本。许多其他样本往往是在交易伙伴之间交往时生成的。例如,市场数据是在出售时生成的,或在(实际的或潜在的)购买者之间调查生成的。因特网(例如,网上拍卖)也是数据来源。

从调查统计学家和调查数据使用者的观点来看,样本调查方面存在着大量的文献。前者讨论如何从总体进行抽样以及从各种不同抽样设计中所得到的结果,而后者讨论利用各种抽样设计收集调查数据时所产生的估计和推论。关键问题

是,样本如何更好地代表总体。本章以一种介绍方式来讨论文献的这两个方面。其他一些细节将在第 24 章给出。

3.2.1 调查数据的特性

观测数据这一术语通常意指,在没有任何企图控制索要抽样数据的特征下,通过抽取对象的相关总体而收集到的调查数据。设 t 表示时间下标,设 $\mathbf{w}$ 表示关注变量的集合。在当前背景下,t 可以是一个时间点或时间区间。设 S_t 表示源于总体概率分布 $F(\mathbf{w}_t|\boldsymbol{\theta}_t)$ 的样本;S_t 是从 $F(\mathbf{w}_t|\boldsymbol{\theta}_t)$ 中抽取的,其中,$\boldsymbol{\theta}$ 表示参数向量。总体应该被看成具有关注特征的点的集合,而且为了简单起见,我们假定概率分布 F 的形式是已知的。简单随机抽样方案允许总体的每一个元素进入样本的概率是相等的。更复杂的抽样方案稍后将加以考虑。

平稳总体(**stationary population**)的抽象概念提供一种有用的基准。如果总体特征的矩都是常值,那么我们可以写成 $\boldsymbol{\theta}_t=\boldsymbol{\theta}$,对于所有 t。这是一个强假设,因为它意味着,总体特征的矩都是时常值。例如,年龄—性别分布应该是常值。更为切合实际地讲,某些总体特征不是常值。为了处理这种可能性,每个总体(的参数)可以被看成从具有常值特征的超总体(**superpopulation**)中抽取的。具体来说,我们认为,每个 $\boldsymbol{\theta}_t$ 是从具有常值(超)参数 $\boldsymbol{\theta}$ 的概率分布中抽取的。在第 24 章讨论的层次模型方面的文献中,经常出现超总体与超参数术语。如果 $\boldsymbol{\theta}_t$ 具有演化分量,那么便引发另外的复杂性,例如,自始至终依赖于 t,或者逐次值是相互依存的。如同第 13 章和第 26 章所讨论的,利用层次模型,将提供对超参数与子总体特征之间关系进行建模的一种方法。

3.2.2 简单随机样本

作为后面讨论的一个基准,考察简单随机抽样,对于所有的 i,从容量为 N 的总体中抽取单元 i 的概率是 $1/N$,其中,N 很大。把 $\mathbf{w}$ 分割成$[y:\mathbf{x}]$。假定我们的兴趣在于对 y 进行建模,y 是以外生协变量向量 $\mathbf{x}$ 为条件的可能向量取值的结果变量,其联合分布记为 $f_J(y,\ \mathbf{x})$。它能够因式分解成为条件分布 $f_C(y|\mathbf{x},\ \boldsymbol{\theta})$ 与边缘分布 $f_M(\mathbf{x})$ 的积:

$$f_J(y,\ \mathbf{x})=f_C(y|\mathbf{x},\ \boldsymbol{\theta})f_M(\mathbf{x}) \tag{3.1}$$

简单随机抽样(**simple random sampling**)包含从整个总体中均匀抽取的$(y,\ \mathbf{x})$组合。

3.2.3 多阶段调查

一种可供选择的方案是分层多阶段整群抽样(**stratified multistage cluster sampling**),也称为复杂调查(**complex survey**)方法。大范围调查,譬如当前人口调查(CPS)和收入动态面板数据调查(PSID),都采用这一方法。24.2 节对 CPS 的体系提供额外的详细内容。

复杂调查设计拥有许多优点。因为该方法减少地理差异,所以它具有更高成

本且有效，而且可能是以更彻底的形式抽取某些子总体。例如，对较小的子总体过度抽样会表现出某些相关的特征，这是可行的，然而对总体的随机样本抽样会产生太少的观测值，不能支持可靠的结果。其缺点是，分层抽样将减少个体间的变异，这在本质上有助于给出较高的准确性。

样本调查文献关注于多阶段调查(**multistage surveys**)，它把总体按次序分割成以下类别：

1. 层(**strata**)：把总体彻底分割为互补相交的一些子总体。
2. 初级抽样单元(**primary sampling units**)(PSUs)：对层分割成互不相交的子集。
3. 第二级抽样单元(**secondary sampling units**)(SSUs)：对 PSU 分割成一些子单元，可以依次分割下去，等等。
4. 最终抽样单元(**ultimate sampling unit**)(USU)：选择最终单元进行采访，它可以是一个住户或者一些住户的集体(段)。

举一个事例，层可以是一个国家的各个不同的州或省，PSU 可以是一个州或省内的地区，而 USU 可以是在相同邻域中形成的小住户群。

通常所有层都要进行调查，例如，所有的州都将肯定进入样本中。但是，并不是全部的 PSU 及其划分被调查到，而且它们以不同比率被抽样。在两阶段抽样(**two-stage sampling**)中，被调查的 PSU 是以随机方式抽取的，而 USU 则是从选取上的 PSU 中以随机方式抽取。在多阶段抽样(**multistage sampling**)中，中间抽样单元譬如 SSU 也会出现。

这些抽样方法的一个结果是，不同的家庭将以不同概率被抽取为样本。于是，此样本是总体非代表性(unrepresentative)的。许多调查都提供一些抽样权数(**sampling weights**)，目的是与被抽取的概率成反比例，在此情况下，这些权数能够用于获得总体特征的无偏估计量。

例如，由于在相同的小邻域内对许多家庭进行抽样，所以调查数据可能是集聚的。在同一个整群中的观测值可能不是独立的或者相关的，因为它们依赖于能够影响到一个层内的所有观测值的某种可观测的或不可观测的因素。例如，郊区或者由高收入家庭占据着，或者由其偏好的某一方面相对同质的那些家庭所占据。源于这些家庭的数据至少将无条件地趋于相关，尽管这样的相关性在以家庭的可观测特性为条件下是可忽视的。忽略样本观测值之间相关性的统计推断产生的方差估计值，比来自正确公式的那些情况所得到的方差估计值要小。

24.5 节将以比较深入的方式涵盖这些问题。两阶段与多阶段样本潜在地使得标准误差的计算更为复杂。

总之，(1) 在一些层内以各种不同抽样比率所得到的分层，意味着样本是总体的非代表；(2) 与被抽取的概率成反比例的抽样权数，可以用于获得总体特征的无偏估计；(3) 集聚会导致观测值的相关，从而低估了估计量的真实标准误差，除非做出适当的调整。

3.2.4 *有偏样本*

如果随机样本是抽取获得的，那么数据的概率分布与总体分布是相同的。对

随机抽样的某种偏离，引起两者之间的差异(**divergence**)，这称为有偏抽样(**biased sampling**)。数据分布会以依赖于对随机抽样偏离的性质而不同于总体分布。由于从子总体中获取数据是更方便的，或者是出于成本考虑的，所以会发生偏离随机抽样的情况，尽管所获取的数据并不是整个总体的代表。现在，我们以没有背离随机性的情况开始，来考虑这类偏离的几个案例。

外生抽样

如果分析者只基于外生变量 $\mathbf{x}$ 而不是响应变量的集合，将可利用的样本分割成一些子样本，就产生了出自调整数据的外生抽样(**exogenous sampling**)。例如，在对德国医院的一项研究中，盖尔等人(Geil et al.，1997)把数据分割成两种类型：患慢性病的人与没有患慢性病的人。由收入范畴来分类也是普遍的。也许更准确的是，把这样的抽样形式描述成外生子抽样，因为它是通过参照已收集到的样本来执行的。通过性别、健康或社会经济地位进行分割是相当普遍的。在外生抽样的假设下，外生变量的概率分布与 y 是独立的，而且不包括关注的总体参数 $\boldsymbol{\theta}$。因此，人们可以忽略外生变量的边缘分布，并且直接建立在条件分布 $f(y|\mathbf{x}, \boldsymbol{\theta})$的基础上进行估计。当然，该假设可能是错误的，同时观测到的结果变量分布会依赖于所选择的分割变量，这或许与结果是相关的，因而导致对外生抽样的背离。

基于响应抽样

如果个体被样本抽取的概率依赖于由那个个体所做出的响应或选择，就产生了基于响应的抽样(**response-based sampling**)。在这种情况下，样本选择可依据由正在研究的内生变量所定义的规则继续进行。

有三个事例如下：(1) 在研究负收入税或援助有抚养孩子负担的家庭计划(**Aid to Families with Dependent Children，AFDC**)对劳动力供给的影响中，仅对那些低于贫困线的人员进行调查；(2) 在研究对公共运输工具样式选择的决定因素中，仅对使用运输工具的使用者进行调查；(3) 在研究对去娱乐场所游玩人数的决定因素中，调查对象至少包括那些去游玩的人。

较低的调查成本为宁愿使用基于选择的样本而不是简单随机样本提供了重要动机。为了生成足够多的相对很少发生的结果或选择观测值(信息)，需要一个非常大的随机样本，因此，比较便宜的方法是去收集源于那些实际上做出选择的人的样本。

这样做的实践重要性是，总体参数 $\boldsymbol{\theta}$ 的一致估计量不再仅仅利用条件总体密度 $f(y|\mathbf{x})$来完成。抽样方案的影响也必须考虑进去。24.4 节将进一步讨论这个专题。

长度偏倚抽样

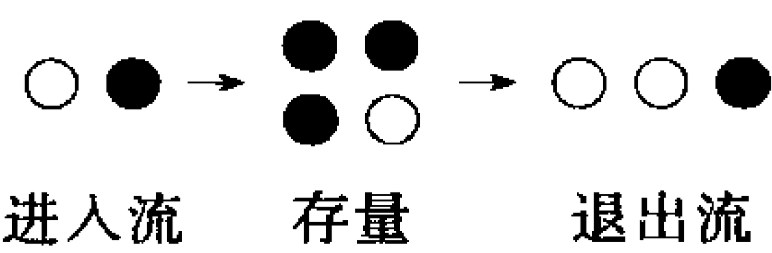
长度偏倚抽样(**length-biased sampling**)阐述，为了对不同总体做出推断，偏倚是如何通过对一个总体进行抽样而引起的。严格地讲，它不足以成为对作为抽取“错误”总体的一个抽样随机性背离的事例。

对过渡进行经济计量研究，就是在其过渡到另一个目的状况 s 之前，对个体 i 处于最初状况 j 所花费的时间进行建模。一个事例是，j 对应于失业，而 s 对应于就业。这类研究所使用的数据可能有几个来源。一个来源是对在特殊日期处于失

业的个体进行抽样；另一个来源是对作为劳动力的那些个体进行抽样，而不管其当前状态如何；第三个来源是对特定时期内成为失业人员或者离开工作岗位的个体进行抽样。每一种抽样方案类型都是基于不同的有关总体概念。在第一种情况下，有关总体是失业个体的存量；第二种情况下总体是劳动力；而第三种情况下总体为满足过渡到就业状况的个体。这个专题将在 18.6 节进一步加以讨论。

假定调查目的是计算失业的平均持续期限测量。这表示随机选取的个体将处于失业时的时间平均长度，如果他或她成为失业者的话。显然，对这个看似简单的问题的解答依赖于样本数据是如何获得的而变化。完成的持续期限的流量分布通常非常不同于存量分布。当我们对存量进行抽样，对于具有较长持续期限的个体而言，处于样本中的概率是比较高的。当我们脱离该状态的流量进行抽样，概率将不依赖于处于该状态所花费的时间。这是著名的基于长度抽样的事例，通过对存量进行抽样而得到的估计值，是随机成为新失业者的失业时段平均长度的有偏估计。

下面这个简单的图示可以解释这一点：



这里，我们用符号“●”表示慢运动者，而符号“○”表示快运动者。假定两种类型在流量中同等地得到表述，而慢运动者在存量中比快运动者停留的时间长。于是，存量总体中的慢运动者的比率较大。最后，推导出总体中的快运动者比例较大。这种推理可推广到其他的异质性类型。

这个事例的要点并不是表明，流量抽样就比存量抽样要好。相反，这要依赖于问题是什么，存量抽样并不会产生有关总体的随机样本。

3.2.5 *由样本选择引起的偏倚*

考察下述问题。研究者对培训的效果测量感兴趣，用 z 表示该效果测量（处理），培训后的工资由 y（结果）表示，给定工人的特征，用 x 表示。如果工人接受培训，那么变量 z 取值为 1，否则取值为 0。对于所有工人来说，观测值（x, D）是可以利用的，但是，只有那些已接受培训（$D=1$）的人才可利用 y。人们喜欢做出有关培训对随机选取的具有已知特征的当前未培训（$D=0$）的工人的培训后工资的平均影响的推断。样本选择(**sample selection**)问题涉及做出这类推断的难点。

曼斯基(Manski, 1995)认为这是一个识别问题，将选择问题正式定义如下：

> 这是对源于随机样本数据的条件概率分布进行识别的问题，条件变量的实现总是可观测到的，但是结果的实现却是删失的。

假定 y 表示要预测的结果，而条件变量用 x 表示。变量 z 表示删失标示变量，如果结果 y 是可观测的，那么 z 取值为 1，否则取值为 0。变量（D, x）总是可观测的，但 y 只有 $D=1$ 时才是可观测的。曼斯基把这称为删失抽样过程(**censored sampling process**)。删失抽样过程不可以识别 $\Pr[y|x]$，正如从下式看到的：

$$\Pr[y|x]=\Pr[y|x, D=1]\Pr[D=1|x]+\Pr[y|x, D=0]\Pr[D=0|x] \quad (3.2)$$

抽样过程能够识别右边四项中的三项，却没有提供关于 $\Pr[y|x,\ D=0]$项的信息。因为：

$$E[y|x]=E[y|x,\ D=1]\cdot\Pr[D=1|x]+E[y|x,\ D=0]\cdot\Pr[D=0|x]$$

无论何时删失概率 $\Pr[D=0|x]$为正，可利用的经验证据对 $E[y|x]$都没有施加约束。因此，删失抽样过程只有对于 $\Pr[y|x,\ D=0]$的某个未知值才能识别 $\Pr[y|x]$。为了认识到关于 $E[y|x]$的一切，必须对 $\Pr[y|x]$施加一些约束。

求解这种问题的可供选择的方法，将在 16.5 节中加以讨论。

3.2.6 *调查数据质量*

样本数据的质量不仅依赖于样本设计和调查工具，而且依赖于调查响应。这种观测值尤其用来表示观测数据。我们考察样本数据的质量受到危机的几种方式。一些问题（比如损耗）连同其他的数据类型也能够产生。这个专题和有偏抽样重叠。

调查无响应问题

正式来讲，调查是自愿的，而且参与的动机依据住户特征或要回答的问题类型不同而系统地变化。个体可以拒绝回答某些问题。如果在拒绝回答的问题和个体特征之间存在着系统关系，那么在考虑无响应(**nonresponse**)之后，就产生了调查代表性的问题。如果可以忽略无响应，而且仅仅利用来自响应的数据完成分析，那么对关注的参数进行估计会受到怎样的影响呢？

调查无响应是前面一节中提及的选择问题的特殊情况。两者都包括有偏样本。为阐述有偏样本是如何导致曲解推断的，考察下述模型：

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}\Bigg|\mathbf{x},\ \mathbf{z}\sim\mathcal{N}\left(\begin{bmatrix} \mathbf{x}'\boldsymbol{\beta} \\ \mathbf{z}'\boldsymbol{\gamma} \end{bmatrix},\begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}\right) \tag{3.3}$$

其中，y_1 表示关注的连续随机变量（比如开支），它依赖于 $\mathbf{x}$，而 y_2 表示潜变量，它测算了调查中的“参与倾向”，依赖于 $\mathbf{z}$。如果 $y_2>0$，那么个体参与；否则，个体不参与。假定变量 $\mathbf{x}$ 与 $\mathbf{z}$ 是外生的。公式允许 y_1 与 y_2 是相关的。

假定我们通过最小二乘法，从参与者提供的数据中估计出 $\boldsymbol{\beta}$。在存在不参与者的情况下，这个估计量是无偏的吗？回答是，如果不参与者是随机的且与关注的变量 y_1 是独立的，那么它就是无偏的；否则，它将是有偏的。

其推理如下：

$$\hat{\boldsymbol{\beta}}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_1$$
$$E[\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}]=E[[\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'E[\mathbf{y}_1-\mathbf{X}\boldsymbol{\beta}|\mathbf{X},\mathbf{Z},\mathbf{y}_2>0]]$$

其中，第一行给出 $\boldsymbol{\beta}$估计值的最小二乘法公式，而第二行说明它是有偏的。如果 $\mathbf{y}_1$ 与 $\mathbf{y}_2$ 是独立的并以 $\mathbf{X}$ 与 $\mathbf{Z}$ 为条件，而且 $\sigma_{12}=0$，那么

$$E[\mathbf{y}_1-\mathbf{X}\boldsymbol{\beta}|\mathbf{X},\mathbf{Z},\mathbf{y}_2>0]=E[\mathbf{y}_1-\mathbf{X}\boldsymbol{\beta}|\mathbf{X},\mathbf{Z}]=\mathbf{0}$$

是无偏的。

缺失数据与误测数据

调查回答者要处理广泛的调查表，他们不一定会回答每个问题，他们即使回答，也可能会故意或偶然地造假。假定调查样本企图获得来自 N 个个体样本的响应向量，记为 $\mathbf{x}_i=(x_{i1},\cdots,x_{iK})$，$i=1,\cdots,N$。现在假定个体没有提供 $\mathbf{x}_i$ 的一个或多个元素的信息，那么整个向量会被丢弃。起因于缺失数据（**missing data**）的第一个问题是，样本量减少。第二更严重的潜在问题是，缺失数据潜在地导致类似于选择偏倚的偏倚。如果数据是以系统方式缺失，那么接下来要分析的样本就不是总体的代表。选择偏倚的形式包括系统的无响应模式。例如，高收入者可能系统地不回答有关收入的问题。相反，如果数据以完全随机形式缺失，那么放弃的不完全观测值将减少准确性，但不会产生偏倚。第 27 章将更深入讨论缺失数据问题及其解决方法。

测量误差（**measurement errors**）在调查响应中是一个普遍性问题。这些问题产生于一系列原因，包括由于粗心大意引起的不正确响应、故意错报、对过去事件不完善的回忆、对问题不正确的解释以及数据过程误差。测量误差更多是由于测量变量最好也不过是相关理论上概念的不完美代表（**proxy**）。这种测量误差的后果是一个主要专题，对此将在第 26 章加以讨论。

样本损耗

在面板数据情形下，调查涉及对一系列个体的重复观测值。在此情况下，我们能够具有：

- 所有时期的全部响应（全部参与）；
- 在第一个时期及后来所有时期都无响应（无参与）；
- 在最初时期响应，但在后来时期没有响应的意义上的部分响应（不完全参与），这种情况称为样本损耗（**sample attrition**）。

样本损耗会产生缺失数据，并且“缺失”的任何非随机模式的存在，将会导致前面提及的样本选择类型问题。这能够解释成为样本选择问题的特殊情况。样本损耗将在 21.8.5 节和 23.5.2 节中简要地加以讨论。

3.2.7 *观测数据的类型*

横截面数据（**cross-section data**）是通过在某些 t，对样本 $\mathcal{S}_t$ 进行观测 $\mathbf{w}$ 而获得的。尽管通常在同一时点上抽样所有住户是行不通的，但横截面数据还是作为对被用于做出有关总体推断的那个总体中子集的每一个元素特征的简略缩影。如果总体是平稳的，那么利用 $\mathcal{S}_t$ 做出有关 $\boldsymbol{\theta}_t$ 的推断，对于 $t'\neq t$ 还是有效的。如果在过去特性和现在特性之间存在显著相依性，就需要纵向数据来确定关注的关系。例如，过去的决策会影响到当前结果；惯性或习惯持续性可以解释当前的购买，但是，如果没有购买历史可以利用，就不能对这类相依性进行建模。

重复横截面数据（**repeated cross-section data**）是通过取自 $F(\mathbf{w}_t|\boldsymbol{\theta}_t)$，$t=1,\cdots,T$ 的一系列独立样本而得到的。因为样本设计并没有企图把相同单元保留在样本中，所以，有关特性方面的动态相依性信息就丢失了。如果总体是平稳的，那么重复横截面数据可通过类似于从恒常总体中进行放回抽样的那种抽样过程来获得。

如果总体是非平稳的，那么重复横截面与依赖于该总体如何随时间而变化的方式有联系。在这种情况下，目标就是对有关基本常值(超)参数进行推断。对重复横截面的分析将在22.7节中加以讨论。

面板或纵向数据(**panel or longitudinal data**)是通过最初选择的样本 S，然后收集一系列时期的观测值而获得到的，$t=1,\cdots,T$。这可以通过访问对象并在同一时间收集现在和过去的数据，或者把对象引进调查中并跟踪它们来获得。这就产生一些列数据向量$\{\mathbf{w}_1,\cdots,\mathbf{w}_T\}$，它们可用于对总体的特性或个体的特殊样本的特性进行推断。每一种情况下的适当方法论是不一样的。如果数据是从非平稳总体中抽取的，那么合适的目标应该是对子总体的(超)参数进行推断。

这些数据类型的一些局限性是明显的。横截面样本与重复横截面通常没有提供对结果中跨期相依性进行建模的适当数据。这类数据仅适合于对静态关系进行建模。相反，纵向数据既适合于对静态关系，又适合于对动态关系进行建模，特别是对纵向数据跨度为足够长的时期建模。

纵向数据并不是没有问题的。第一个问题是面板的代表性。如果总体是非平稳的，那么利用纵向数据对有关总体特性进行推断是相当困难的。为了分析行为动态学，只要可能在面板中保留最初住户，就是引人注目的选项。在实际应用中，纵向数据集合遭受"样本损耗"问题，或许是由于"样本疲劳"(sample fatigue)。这就意味着，调查回答者并没有连续提供对调查表的响应。这导致两个问题：(1) 面板成为非平稳的；(2) 存在着下述危险，即被保留住户不是一般性，并且样本成为总体的非代表。当可利用的数据不是从总体中随机抽取的时候，建立在各种不同数据类型基础上的结果会不同程度地对偏倚有敏感性。由于把个体保留在不同时期的面板中是相当困难的，或者由于某种其他原因，比如位置的改变，个体被"丢失"(删失)，这就产生了"样本疲劳"。有关这些问题，本书稍后将会研究。不过，纵向数据的分析可以提供抽样单元特征的某些方面信息，尽管外推总体特性并不是简单易行的。

3.3 源自社会实验的数据

观测数据和试验数据极为不同，因为实验环境原则上接近于可监督与可控的。这使得改变关注的原因变量成为可能，而把其他协变量固定在可控的设置背景下。相反，观测数据是在非可控环境下生成的，这留下一种公开的可能性——混淆因素的存在将会使得对关注的因果关系进行识别更为困难。例如，当人们试图利用观测数据研究工资—受教育关系时，人们必须接受个体的受教育年数是个体者自身决策过程的结果，而不能把受教育水平看成由假设实验者来设置的。

在社会科学中，实验数据与类似的数据或者来自社会实验(**social experiments**)，本书下面将深入而详细地定义和描述它；或者来自"实验室"实验("laboratory" experiments)，即一个自愿参与者小组处于同现实生活相对应的实验中，模拟经济行为人的行为。

本节提供对社会实验方法论的简要解释、源于社会实验数据的性质、由此产生

的问题，以及对经济计量方法论的争论。

实验方法论的核心特性涉及随机选取的实验小组受限于“处理”(**treatment**)的结果与那些对照(**control**)(比较)组之间的比较。在一个良好的实验中，对对照组与实验(“处理”)组的匹配关系要进行相当仔细的审查，从而避免结果中的潜在偏倚。这样的条件在观测环境中不可能实现，因此会导致对关注的原因参数识别的可能缺失。然而有时，实验条件可以近似地由观测数据来复制。例如，考察邻近的地区或州，其中一个地区执行与另外一个地区不同的最低工资政策，创造了自然实验(**natural experiment**)的条件，来自“处理州”的观测值能够与那些来自“控制州”的观测值进行比较。经济计量学中的自然实验的数据结构也是引人关注的问题。

社会实验涉及包含一系列受试者的经济环境中的外生变化，受试者被分成接受实验处理的子集与另外一个作为对照组的子集。与观测研究——外生因素与内生因素的变动经常混淆在一起——相比，设计良好的社会实验目的是隔离处理变量的作用。在一些实验设计中，可能不存在明显的对照组(**control group**)，却可以利用处理的变化水平，在此情况下，原则上估计实验结果的整个响应面(**response surface**)是可能的。

社会实验的主要目的是，估计实际或潜在的社会项目的效应。2.7 节的潜在结果模型提供了对社会实验效应进行建模的相关影响。几种可供选择的测量效应的方法已经提出来，这些将在项目评估章(第 25 章)中加以讨论。

伯特莱斯(Burtless，1995)已概述过社会实验的情况，又注意到一些潜在的局限性。在同类文章中，赫克曼和史密斯(Heckman and Smith，1995)关注可以执行的实际社会实验的局限性。本节后面的讨论明显借用这些论文的观点。

3.3.1 *社会实验的重要特性*

社会实验是由下述政策问题引发的，即受试者对从未执行的政策类型是如何反应的，因此，对此没有观测相应数据社会实验的思想是，去征募一个自愿参与者小组，将其中一些参与者随机地指派到处理组，而将其余的参与者指派到对照组。在那些受限于政策变化的处理组的响应与那些没有政策变化的对照组的情况之间的差异，就是政策的估计效应。标准实验设计被系统地描述成图 3.1。

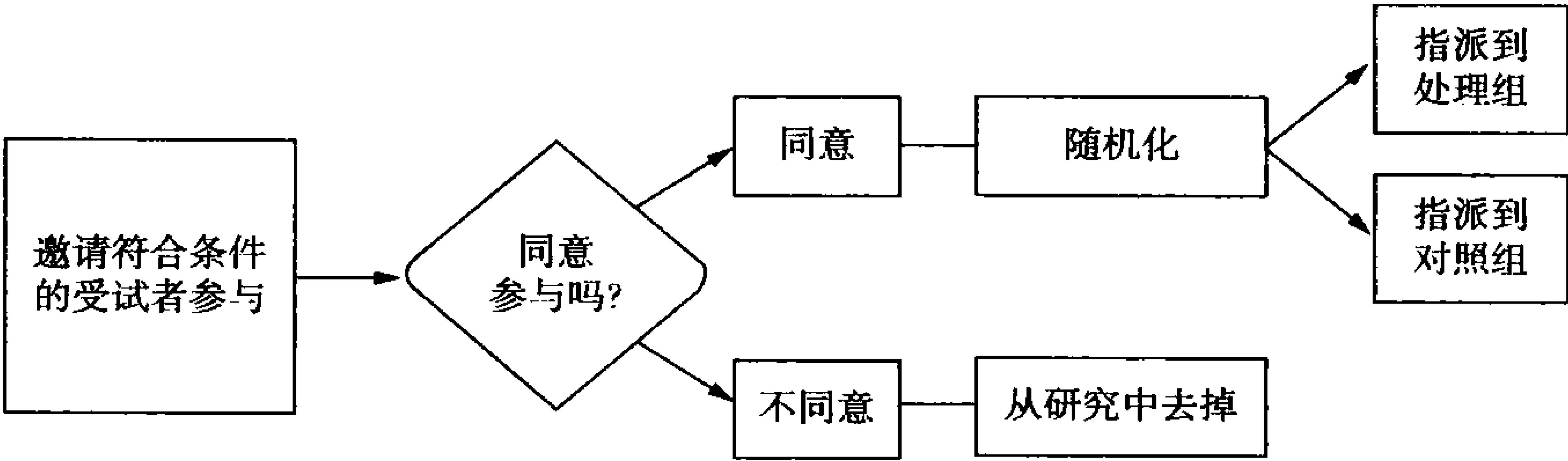


图 3.1 带有随机指派的社会实验

术语“实验”意指接受处理的组，“对照”(“控制”)意指未接受处理的组，而“随机指派”(**random assignment**)意指指派个体到上述两个组的过程。

统计学中的“随机化试验”是由 R. A. 费希尔(R. A. Fisher，1928)同其合作者

一起引进的。典型的农业实验是由下述实验构成的，即一种新处理，比如肥料被应用到一块随机选取的植物生长地上，那么把其响应与那些对照组的植物进行比较，这对于实验中所有相关的方面都类似，只是没有给出实验处理。如果实验组与对照组之间的所有其他差异的效应都被剔除掉，那么在两个响应集合之间的估计差异被认为是由处理所造成的。在最简单的情况下，人们能够集中精力比较处理组的均值结果与未处理组的均值结果。

尽管在农业和生物医学科学中，随机化试验方法论具有悠久的历史，但在经济学和社会科学中，它却是崭新的。研究那些没有观测数据的政策变化的响应是引人注目的，或许因为关注的政策变化从未被执行过。随机化实验还允许政策变化和参数的变动比其在观测数据中存在的变动要大许多，因而，对政策变化的响应进行识别与研究就比较容易。在许多情况下，社会实验可以彻底检验从未执行的政策，因此，观测数据在其潜在影响中完全可以保持沉默。

社会实验除了在美国以外还是相当少的，部分原因在于它们实施起来费用昂贵。在美国，一系列的这种实验始于 20 世纪 70 年代早期。表 3.1 概述某些相对著名事例的特性；有关更广泛的内容，参见伯特莱斯(Burtless, 1995)。

表 3.1 一些选出的社会实验的特点

实　验	实验处理	目标总体
兰德健康保险实验(RHIE)，1974～1982 年	健康保险计划和不同的最大支出费用	低水平及中等水平收入人员与家庭
负税收(NIT)，1968～1978 年	带有可选择收入保证和税率的 NIT 计划	未成年户主的低水平及中等水平收入人员与家庭
职业培训关系法(JTPA)，1986～1994 年	在 JTPA 融资下，寻找职业资助、工作培训、课堂培训	失学青年人和贫困成年人

通常，实验可能产生横截面数据，或是纵向数据，尽管出于成本考虑，时常会将时间维度限制在观测数据中的某一个水平以下。当一个试验持续几年，并且是多阶段的以及/或拥有一些地理场所，如同 RHIE 情况，建立在不完全数据基础上的期间分析并不罕见[纽豪斯等人(Newhouse et al., 1993)]。

3.3.2 *社会试验的优点*

伯特莱斯(Burtless, 1995)非常明晰地综述了社会实验的优点。其重要的优点源于随机化试验，这可以消除项目参与中观测到的特征和未观测到的特征之间的相关性。因此，如果没有混淆偏倚，即使人们不能对混淆变量加以控制，处理对已处理组和对照组之间的差异的贡献就能得到估计。处理变量和起混淆作用的变量之间存在的相关性经常困扰着观测研究，同时使得因果推断复杂起来。与之相比，在理想环境下进行的实验研究，可以产生已处理组和未处理组的结果平均差异的一致估计值，而在计算上又没有更多的复杂性。

然而，如果结果依赖于处理和其他可观测的因素，那么对可观测因素加以控制，通常会改进估计效果的准确性。

显然，可以利用观测的数据，但实验数据的产生和使用具有很大的感染力，因

为它提供了对政策变量外生化(exogenizing)的可能性,同时,处理的随机化能潜在地导致统计分析在很大程度上的简化。建立在观测数据基础上的结论经常具有一般性,因为它们是建立在来自总体的非随机样本基础上——选择偏倚问题。一个例子就是前面提及的 RHIE 研究,它主要专注的内容是对健康服务需求的价格响应。健康保险的可用性会影响到健康服务的使用者价格以及对它的使用。一个重要的政策问题是,对健康服务过度使用的程度是由已补助健康保险引起的。当然,人们能使用观测数据对健康服务和保险水平之间的关系进行建模。然而,这类分析受到下述批评,即健康保险水平不应该被处理成为外生的。理论上的分析表明,对健康保险的需求与健康保险可以联合确定,因此,因果关系不是单方面的。这一事实会潜在地使得对健康保险识别很困难。把健康保险处理成外生偏倚的价格响应的估计值。然而,在实验背景下,参与的住户/家庭被指派一个保险政策,使它作为外生变量。于是,保险的作用是可识别的。一旦关注的重要变量被外生化,因果关系的方向就变得清楚,而且处理效果能得到清楚的研究。进一步地,如果实验没有我们下面将提及的一些问题,就会大大简化有关的统计分析,在调查数据中这种分析常常是必需的。

3.3.3 *社会实验的局限性*

非人类方法论——起初发展并应用到非人类受试者——应用到人类受试者,在文献中产生了广泛的讨论。特别地,可参见赫克曼和史密斯(Heckman and Smith, 1995),他们认为,许多社会实验可能遭受到应用于观测研究上的局限性。这些问题涉及诸如实验方法论与观测方法论的比较,以及使用于人类受试者时存在的内在偏倚和问题。后面几章会详细地讨论这几个问题,但是此处仅提供概览。

社会实验实施起来成本非常高。有时,社会实验或许经常不对应于“纯洁的”随机化试验。因此,出自这类实验的结果并不总是清晰明确且容易解释的,或者是没有偏倚的。如果处理变量具有许多可供选择的关注设置,或者如果外推是一个主要目的,那么必须搜集非常庞大的样本来确保充分的数据变异,且准确地对处理变化的效应进行估计。在这种情况下,实验成本也会增大。如果成本因素阻碍了大量实验,那么它与观测研究有关的效用可能是不可靠的;参见豪斯曼和怀斯(Hausman and Wise, 1985)的《社会实验》中由罗森(Rose)和斯塔福德(Stafford)所撰写的文章。

不幸的是,某些社会实验的设计是有缺陷的。豪斯曼和怀斯(Hausman and Wise, 1985)曾讨论,源于新泽西州负收入税实验的数据受限于内生分层,他们进行了如下描述:

> ……实验的原因是通过随机化,去掉处理变量和正在研究的其他响应变量的决定因素之间的相关性。然而,在收入生活费实验中,实验样本的选取是部分建立在因变量基础上的,并且指派处理及对照组也是部分建立在因变量基础上的。通常,适合于选择的组——建立在家庭状况、种族、住户年龄等基础上——是在收入(或其他变量)基础上进行分成的,而人员则是从每一个层内选取出来的[豪斯曼和怀斯(Hausman and Wise,1985,第 190~191 页)]。

作者得出结论,在存在内生分层的情况下,处理效应的无偏估计并不能简单推导出来。不幸的是,完全随机化的试验的成本极高且可能并不可行,其中,从总体中随机抽取的实验组内的处理指派是与收入独立的。

存在损坏随机化实验理想简单性的几个问题。第一,如果实验场所是随机选取的,就会需要那个场所的提供者和潜在参与者的合作。如果这不是现成的,那么可获得这类合作的可供选择的场所将作为替代的,因此危及随机指派原理;参见霍茨(Hotz, 1992)。

第二,是关于样本选择问题,由于参与是自愿的,所以这是与之紧密相关的。由于一些道德原因,存在许多不能简单实施的实验(例如,随机指派学生受教育年数)。与医学实验能够达到双盲治疗方案的黄金标准不同,在社会实验中,实验者和受试者知道他们是否处于处理组或对照组。进一步地,对照组可能获得出自可供选择来源的处理(例如,培训)。如果所做出的参与决策与 **x** 或 ε 是不相关的,那么实验数据会得到简化。

第三个问题是由受试者在实验开始之后从实验中产生的样本损耗。即使最初样本是随机的,非随机损耗的效应可能会产生类似于面板中损耗偏倚的问题。最后,存在霍桑效应(**Hawthorne effect**)的问题。这一术语起源于社会心理学研究,该研究是由哈佛商业管理研究生院与西方电力公司管理部门在芝加哥的霍桑工厂从1926年到1932年所开展的。不像无生命物体,人类受试者尽管在实验中是参与者,但可以改变或适应他们的行为。在这种情况下,在实验中观测到的响应的变化不能被认为是仅由处理造成。

赫克曼和史密斯(Heckman and Smith, 1995)提到了在实行随机化处理中的其他几个困难。由社会实验的管理涉及政府机构,存在着潜在的偏倚。在试验正常运作下,如果指派引进了实验参与者与参与者之间的系统差异,就产生了随机化偏倚(**randomization bias**)。豪斯曼和史密斯用文章证明了真实实验存在这类偏倚的可能性。另外一种偏倚类型称为替换偏倚(**substitution bias**),当对照组可接受某种形式的处理并用来代替实验处理时,就产生了这种替换偏倚。最后,社会试验的分析必然具有局部均衡性质。人们不能以可靠方式把处理效应外推到整个总体,因为当涉及总体时,其余条件不变(**ceteris paribus**)的假设将不再成立。

特别地,核心问题是,人们是否能够把从实验中得到的结果充分地外推到总体上。如果实验在一个小规模内作为探索项目来实施,但计划是要预测更广泛地应用该政策的效应,那么其明显的局限性是,实测探索项目将不能包括处理的较广泛的效果。比较广泛应用的处理会改变经济环境,这将充分证明,在局部均衡背景下的预测是错误的。因此,这种处理将不会像它所模拟的真实政策那样。

总之,社会实验原则上能够产生数据,运用这种数据比观测数据更容易就原因与效应进行分析和认识。这一目标是否实现将依赖于实验设计。不好的实验设计会导致统计复杂,并影响到结论的预测。社会实验本质上不同于生物学和农业上的那些实验,因为人类受试者和处理提供者既是积极的又是有远见的个体,他们具有个人偏好,而不是标准治疗方案的消极提供者,也不是自愿的接受随机指派处理的接受者。

3.4 源自自然实验的数据

然而，有时研究者可以利用源于“自然实验”(**natural experiment**)的数据。当总体的一个子集受限于外生变量的变化，或许作为政策变动的结果，这常常受限于内生变化，就产生了自然实验。原则上，变化的来源是很好理解的。

在微观经济学中，利用自然实验的思想有两种广泛使用的方式。为了具体起见，考察简单回归模型：

$$y=\beta_1+\beta_2 x+u \tag{3.4}$$

其中，x 表示与 u 相关的内生处理变量。

假定存在一个外生干预，它会改变 x。这种外在干预的事例包括行政法规、非预期法律、自然事件(例如，双胞胎出生)、有关天气的变动以及地理变化；参见表 3.2 中事例。外生干预创造了一种通过比较干预前和干预后影响的行为或者干预后非影响组的行为来估计其效果的机会。也就是说，“自然”比较小组是由推进 β_2 估计的事件生成的。因为 x 可看成外生的，所以估计得到简化。

表 3.2 一些选出的自然试验的特性

实　验	处理研究	文　献
具有不同受教育水平的双胞胎的结果	尽管受教育与年龄之间相关，但受教育回报上有差异	阿申费尔特和克鲁格(Ashenfelter and Krueger, 1914)
加拿大萨斯喀彻温省国家健康保险转变到 NHI 及后来持续多年的其他形式	建立在具有 NHI 与没有 NHI 的省份比较基础之上的 NHI 的劳动力市场影响	格鲁伯和汉拉迪(Gruber and Hanratty, 1995)
新泽西州提高最低工资而邻近的宾夕法尼亚州没有变化	最低工资对就业的影响	卡德和克鲁格(Card and Krueger, 1994)

自然实验可以帮助推断的第二种方法，是通过生成自然工具变量来进行的。假定 z 表示一个变量，它与 x 是相关的，或者在原因形式上与 x 有关，且与 u 无关。于是，β_2 的工具变量(**instrumental variable**)估计量可用样本协方差形式表示，即：

$$\hat{\beta}_2=\frac{\mathrm{Cov}[z,y]}{\mathrm{Cov}[z,x]} \tag{3.5}$$

(参见 4.8.5 节。)在观测数据背景下，要找出具有正确性质的工具变量可能很困难，但是，在有利的自然实验中，却可以自然地产生工具变量。于是，估计得以简化。我们将在下一节考察第一种情况；第 25 章将讨论自然生成工具这一专题。

3.4.1 *自然外生干预*

收集这类数据并不昂贵，而且这类数据允许研究者去评估隔离中的某一特定因素的作用，如同可控实验一样，因为“自然”对其他并不直接关注的因素来说，提供恒定变异的贡献。这种自然实验是引人注目的，因为它们花费不多，且在现实世界背景下就可生成处理组和对照组。一个自然实验能否支持令人信服的推断，部

分地依赖于支持自然干预是否真正是外生的，它的影响是否足以达到可测量的程度以及是否存在良好的处理组和对照组。例如，正因为变动是通过立法而产生的，所以并不意味着它是一个外生的干预。然而，在适当情况下，对这类数据集不失时机地利用，能够产生有价值的实证观点。

建立在自然实验基础上的研究探索具有几个潜在的局限性，在任何一项给定研究中，其重要性只能通过仔细地考虑有关理论、事实以及制度背景才会得到评估。沿着坎贝尔(Campbell, 1969)和迈耶(Meyer, 1995)的线索，可以把局限性分为：影响研究内部有效性的影响(也就是说，对从研究中推出的政策影响进行推断)与影响研究外部性的影响(也就是说，把结论推广到总体中的其他元素上)。

利用下面将简要描述的以及将在第 25 章详细讨论的回归方法，通过比较从干预前与数据干预后中推导出的结论，来考察政策变动的调查研究。在任何研究中，都存在省略变量——在政策变动与其影响之间的时间区间上还会变化的变量。所抽取的个体特征，诸如年龄、健康状况以及他们的真实经济环境或预期的经济环境，可能也会有变化。这些省略因素将直接影响到政策变动的影响。结果能否推广到总体中其他元素上，将依赖于缺少由非随机抽样而引起的偏倚、政策变动和其背景之间交互作用效应的显著存在，以及缺少引起影响从一种情况到另一种情况改变的历史因素。当然，这些考虑对于来自自然实验的数据而言不是唯一的；然而，其意义是，后者不一定就不受这些问题的困扰。

3.4.2 *差异中的差分*

一种简单的回归方法是建立在对政策干预之前与之后的同一个组结果的比较基础之上。例如，考察：

$$y_{it}=\alpha+\beta D_t+\varepsilon_{it}, \quad i=1,\cdots,N,\ t=0,1$$

其中，$D_t=1$ 表示处于第 1 时期(干预后)，$D_t=0$ 表示处于第 0 时期(干预前)，而 y_{it} 测量结果从合并数据中估计出来的回归，将产生政策影响参数 β 的估计值。很容易证明，这等于干预前与干预后结果的平均差：

$$\begin{aligned}\hat{\beta} &= N^{-1}\sum_i(y_{i1}-y_{i0})\\ &= \bar{y}_1-\bar{y}_0\end{aligned}$$

一个组在设计之前与之后要做出强的假设：该组对于不同的时间保持可比性。这就要求 β 具有可识别性。例如，如果我们允许 α 在两个时期之间变化，那么 β 就不再是可识别的了。α 的变化会与政策影响混淆。

对前面设计加以改进的一种方法是包括一个附加的未处理比较组，也就是说，它未受到政策影响，而其数据在两个时期都是可利用的。利用迈耶(Meyer, 1995)的记号，现在有关的回归是：

$$y_{it}^j=\alpha+\alpha_1 D_t+\alpha^1 D^j+\beta D_t^j+\varepsilon_{it}^j, \quad i=1,\cdots,N,\ t=0,1$$

其中，j 表示组的下标。如果 $j=1$，那么 $D^j=1$；否则，$D^j=0$。如果 $j=1$ 且$t=1$，那么 $D_t^j=1$；否则，$D_t^j=0$。ε 表示具有零均值和常值方差的误差项。该方程没有包含协变量，但是可以添加它们，而那些不变化的项已经归入 α 中。这种关系蕴含着，

对于已处理组来说，我们有干预前的：

$$y_{i0}^{1}=\alpha+\alpha^{1}D^{1}+\varepsilon_{i0}^{1}$$

与干预后的：

$$y_{i1}^{1}=\alpha+\alpha_{1}+\alpha^{1}D^{1}+\beta+\varepsilon_{i1}^{1}$$

因此，其影响是：

$$y_{i1}^{1}-y_{i0}^{1}=\alpha_{1}+\beta+\varepsilon_{i1}^{1}-\varepsilon_{i0}^{1} \tag{3.6}$$

对于未处理组而言，其相应的方程是：

$$y_{i0}^{0}=\alpha+\varepsilon_{i0}^{0}$$

$$y_{i1}^{0}=\alpha+\alpha_{1}+\varepsilon_{i1}^{0}$$

从而，其差为：

$$y_{i1}^{1}-y_{i0}^{0}=\alpha_{1}+\varepsilon_{i1}^{0}-\varepsilon_{i0}^{0} \tag{3.7}$$

这两个一阶差分方程都包含第 1 时期的特殊效应 α_1，这能通过对方程(3.6)与方程(3.7)取差分而得以剔除：

$$(y_{i1}^{1}-y_{i0}^{1})-(y_{i1}^{0}-y_{i0}^{0})=\beta+(\varepsilon_{i1}^{1}-\varepsilon_{i0}^{1})-(\varepsilon_{i1}^{0}-\varepsilon_{i0}^{0}) \tag{3.8}$$

假定 $\mathrm{E}[(\varepsilon_{i1}^{1}-\varepsilon_{i0}^{1})-(\varepsilon_{i1}^{0}-\varepsilon_{i0}^{0})]=0$，我们能通过$(y_{i1}^{1}-y_{i0}^{1})-(y_{i1}^{0}-y_{i0}^{0})$的样本平均来获得 β 的无偏估计值。这一方法使用了差异中差分(**differences in differences**)。如果存在时变协变量，那么它们包含在有关方程中，而且其差分将出现在回归方程(3.8)之中。

为了简单起见，我们的分析忽略了处理组与对照组的特征分布之间存在的观测差异。如果是这样的话，那么这类差异就必须加以控制。其标准解决方法是回归中包含这类控制变量。

建立在自然实验基础上的一个研究事例是，阿申费尔特和克鲁格(Ashenfelter and Krueger, 1994)的研究。他们通过将同一双胞胎的工资率与其不同的受教育水平加以对比来估计受教育回报。在此情况下，实施常规的实验是不可行的，实验中的个体被外生地指派了各种不同的受教育水平。不过，某个实验类型的控制是需要的。正如作者解释的：

> 我们的目标是，确定我们在受教育和工资率之间所观测到的相关性，而不是来自受教育和工人能力或者其他特征之间的相关性。我们这样做利用了下述事实：单一受精卵双胞胎一般是相同的，并且具有相似的家庭背景。

双胞胎数据被用作一系列其他经济计量研究的基础[罗森茨韦格和沃尔平(Rosenzweig and Wolpin, 1980)；布罗纳斯和格罗格(Bronars and Grogger, 1994)]。由于总体中双胞胎概率并不高，一个重要问题是收集充分大的代表性样本，考虑某种无响应。这类数据的一个来源是人口普查，另一个来源是在美国举行的"双胞胎节日"。阿申费尔特和克鲁格(Ashenfelter and Krueger, 1994，第 1 158 页)曾经报告了他们从第 16 届双胞胎节日，即 1991 年 8 月在俄亥俄州的特维斯伯格举行的年度节日，通过访问所获得的数据知，这个节日是世界上最大的双胞胎、

三胞胎以及四胞胎聚会。

利用双胞胎数据的好处是，存在的既出自可观测因素又出自不可观测因素的共同效应，能够通过对双胞胎结果之间的差异进行建模而去掉。例如，阿申费尔特和克鲁格曾估计双胞胎中的老大和老二之间工资率对数差异的回归模型。第一个差异运算可去掉年龄、性别、民族地位等影响。剩下的解释变量是受教育水平——作为主要关注的变量——和诸如职位年限与婚姻状况变量之间的差别。

3.4.3 *通过自然实验进行识别*

自然实验学派对经济计量时间具有十分有益的影响。通过利用半实验数据的机会以及利用诸如第 2 章 POM 的建模框架，经济计量时间在观测数据与实验数据之间的空白处架起一座桥梁。源于 SEM 框架的参数识别概念，可扩展到从政策观点来看非常有趣的对测量的识别。利用源自自然实验数据的主要优点是，关注的政策变量可以被处理成外生的。然而，在利用源自自然实验的数据中，如同在社会实验情况一样，对对照组的选择，在确定结论的可靠性方面起着极其重要的作用。影响社会实践的几个潜在问题，譬如选择性和损耗偏倚，在自然实验情况下仍旧是一些潜在问题。引起关注的政治问题的一个子集，会在自然实验框架内加以分析。实验仅对总体中的一小部分可以使用（应用），而且它的发生条件不会很容易地自我重复。22.6 节给出的一个事例在差异中的差分背景下阐明这一点。

3.5 应用研究

尽管存在着对微观数据的多种数字和类型的应用，但建立良好的数据库可以支持大量的研究。我们提供一些在美国著名的数据库中非常少的一部分目录。对于进一步详细内容，参见这些数据的各自网站或者下面提及的数据信息中心，其中的一些允许你直接下载数据。

3.5.1 *微观数据的某些来源*

收入动态面板研究（Panel Study in Income Dynamics，PSID） PSID 从 1968 年开始执行全国性调查，其调查研究中心建立在密歇根大学。目前，它涵盖 40 000 多人，并且收集经济和人口数据。这些数据用于支持相当广泛的微观经济计量分析。布朗、邓肯和斯塔福德（Brown，Duncan and Stafford，1996）曾概述 PSID 数据的最新发展。

当前人口调查（Current Population Survey，CPS） 这是一个对 50 000 个住户（家庭）的每月国家调查，它提供了劳动力特征的信息。这种调查已经执行了 50 年以上。对样本的主要修改来自每一次人口普查。有关这个调查的其他详细内容，参见 24.2 节，它已成为许多联邦政府统计工资和失业的基础。它还特别是支持劳动力市场大量研究的主要微观数据来源。该调查在 1994 年被重新设计[波利文卡（Polivka，1996）]。

全国纵向调查(National Longitudinal Survey, NLS) NLS具有四个最初分组:NLS老年男人、NSL青年男人、NLS成年妇女、NLS青年妇女。每一个最初分组是对大致5 000个个体的国家每年调查,这些个体从20世纪60年代中期开始重复访问。调查要收集每一个回答者的工作经历、教育、培训、家庭收入、家庭组成、婚姻状况以及健康状况。有关年龄、性别等补充数据都可以利用。

全国纵向青年调查(National Longitudinal Surveys of Youth, NLSY) NLSY是国家每年对12 686个青年男子和青年女子的调查,他们的年龄从14岁到22岁,第1次调查是从1979年开始的。它包括三个子样本。该数据为研究青年人大样本的生活方式提供了机会,而这些青年人是美国男人和妇女的代表,他们出生在20世纪50年代后期和60年代早期。第二次NLSY开始于1997年。

收入和项目参与调查(Survey of Income and Program Participation, SIPP) SIPP是每月对大约8 000个家庭单元进行的纵向调查。它涵盖了收入来源、参与政府津贴项目、这些项目之间的相关性,以及个人参加职业市场期限。它在每一个日历年的开始引进新的面板数据的一种多重面板调查。SIPP的第一个面板数据开始于1983年10月。与CPS相比,SIPP具有较少的就业者和较多的失业者。

健康和退休研究(Health and Retirement Study, HRS) HRS是对国家的纵向研究。其基础是由1992年开始、持续12年、对7 600个家庭成员(回答者年龄从51岁到61岁)的采访数据构成的,其中,后续访问每两年进行一次。数据包括经济财产、人口特征以及健康信息。

世界银行生活标准测量研究(World Bank's Living Standards Measurement Study, LSMS) 世界银行LSMS家庭调查搜集许多发展中国家的数据,这些数据是关于"家庭健康的许多难度可用于评估家庭的福利、认识家庭行为以及估计各种政府政策对人们生活条件的效应"。运用这些数据的一些例子可在迪顿(Deaton, 1997)以及经济发展文献中找到。格罗什和格利(Grosh and Glewwe, 1998)曾经概述数据的性质,并提供了利用它们进行研究的参考文献。

数据交换(Data ClearingHouses) 政治和社会研究大学联盟(The Interuniversity Consortium for Political and Social Research, ICPSR)提供许多数据集,包括PSID、CPS、NLS、SIPP、国家医疗费用支出调查(NMES)以及其他数据。美国劳动力统计局掌握着CPS与NLS。美国人口普查局掌握着SIPP。美国国家健康统计中心提供许多健康数据集。通往欧洲数据档案的有用途径是欧洲社会科学数据档案委员会(CESSDA),它提供对几个欧洲国家数据档案的链接。

期刊数据档案(Journal Data Archives) 就许多目的而言,诸如为课堂教学工作复制已出版的结果,你可以从期刊档案(**journal archives**)中获得数据。特别地,两个档案具有利用因特网浏览器进行上传与下载数据的建立良好的程序。《商业和经济统计学杂志》(*Journal of Business and Economic Statistics*)拥有可用于那本期刊中大部分但不是全部已出版的文章数据。《应用经济计量学杂志》(*Journal of Applied Econometrics*)数据档案也是以类似方式组织的,并且包括与从1994年开始出版的大部分文章有关的数据。

3.5.2 *处理微观数据*

微观经济数据集趋向于十分庞大的集合。容量为成百上千的样本是普通的，甚至那些容量为数十万的样本也并不奇怪。关注结果的分布经常是非正态的，其部分原因常常是处理离散数据，譬如二值结果，或者处理具有有限变异的数据，譬如比例或份额，或者处理截取或删失连续结果的数据。处置大量的非正态数据，会产生对数据重要特性进行概括和报告的一些问题。

3.5.3 *数据准备*

微观经济计量分析的最基本的特征是：有关使样本最终应用于经济计量研究的过程，可能具有悠久的历史。重要的是，在对数据"整理"的过程中，研究者要准确地利用大量事实来做出决策和选择。让我们考虑一些特定的事例。

样本调查数据最普遍的特征之一是无响应（**nonresponse**）或者部分响应。无响应问题已经讨论过。部分响应通常意味着，调查问题表中的一部分没有得到回答，如果这意味着所需要的信息有一部分是不可以利用的，那么不确定的观测值就要被删除，这称为逐表删除[1]（listwise deletion）。如果这种问题以显著数量出现，就应该正确地加以分析并且报告，因为它会导致非代表性样本以及估计偏倚。这一问题将在第27章加以分析。例如，考虑家庭调查中那些高收入家庭没有做出响应的问题，会产生对这些家庭未充分代表的样本。因此，最终效果与那些存在完全响应但样本是非代表性的情形并无不同。

第二个问题是报告数据中的测量误差（**measurement error**）。微观经济数据一般具有噪声。测量误差的范围、类型以及严重性，均依赖于调查是横截面还是面板类型、对调查做出响应的个体以及有关所要寻找信息的变量，例如，来自面板调查的自报告收入数据，被认为具有很强的序列相关测量误差。相反，所报告的开支费用数目通常被认为具有较小的测量误差。迪顿（Deaton, 1997）调查了特别参考世界银行"生活标准测量调查"的测量误差来源，尽管所产生的几个问题具有广泛的关联性。来源于测量误差的偏倚依赖于用变换形式对数据所做的改变（例如，一阶差分）以及所使用的估计量。因此，为了对有关来自测量误差的严重性做出有价值的陈述，人们必须对定义良好的模型加以分析。后面几章将给出在特定背景下测量误差影响的一些事例。

3.5.4 *检查数据*

在极多数据的集合中，很容易产生由键盘录入和编码错误引起的错误数据。因此，人们应该使用一些基本检查，以便揭示存在的问题。人们在对数据分析之前，通过审查一些描述统计学来检查数据。下面一些技术是有用的。第一，运用概括统计量（最小值、最大值、均值以及中位数）来确保数据位于正常区间与正常尺度上。例如，类型变量应该是介于0～1之间。计数则应该大于或等于0。有时候，缺

[1] 又称为单举法剔除。——译者注

失数据标记为－999，或者为某些其他整数，因此，一定不要把这些处理成数据。第二，人们应该知道变动是以分数尺度还是以百分比尺度衡量。第三，利用盒须图(box and whisker)来识别有问题的观测值。例如，利用盒须图，研究者发现一个具有负人口增长的国家(归因于战争)，而另外一个国家所报告的投资大于GDP(因为外国援助被从GDP中排除掉)。在继续进行估计前，检查观测值还可以建议，对于适合建立一个特殊数据集的特性，进行正态化变换与或分布假设。第三，筛选数据(screening data)建议对适当的数据进行变换。例如，画盒须图与直方图，能建议哪些变量通过对数变换或幂变换会比较适合于建模。就某些目的而言，比如使用非线性估计量、改变变量大小以使它们具有大致相同的尺度，这是人们所希望的。概括统计量可用于检查变量的均值、方差以及协方差，从而显示正确尺度。

3.5.5 *展现描述统计量*

因为微观数据集通常是巨大的，极其重要的是，要向读者提供用于描述每一个变量的统计量的最初表格，它通常包括均值、标准差、最小值以及最大值。在一些情况下，出乎意料的大值或者小值，会揭示出全部记录误差或并入了不正确数据点错误的存在。通常，双向散点图不是有用的，但类别变量(又称属性变量)列表(列联表)却是有益的。对于离散变量来说，直方图是有用的；而对于连续变量来说，密度图则提供有用信息。

3.6 文献注释

3.2 迪顿(Deaton，1997)曾经特别提供关于发展中经济的抽样调查引论。第24章将提供复杂调查的几个特定参考文献。贝克迪等人(Becketti et al.，1988)曾研究PSID代表性问题的重要性。

3.3 由豪斯曼和怀斯(Hausman and Wise，1985)主编的文集包括几篇有关个人社会实验的论文，其中包含RHIE、NIT以及分时电价实验(Time-of-Use pricing experiment)。一些研究质疑实验数据的有用性，而且对妨碍得出结论的实验设计方面的缺点存在广泛讨论。伯特莱斯(Burtless，1995)以及赫克曼和史密斯(Heckman and Smith，1995)的两篇杰出论文讨论了社会实验与观测数据的优缺点。

3.4 《商业和经济统计学杂志》(*Journal of Business and Economic Statistics*，1995)的特定专刊，发表了运用准实验和自然实验的一系列论文。文集包括迈耶综述了源于自然实验数据的经济计量研究方法论及其问题的论文。他还遵循自然变动方面一系列有价值的指导路线，这些都部分地建立在坎贝尔(Campbell，1969)的研究基础上，金和辛安尔(Kim and Singal，1993)利用航空公司并购，研究市场集中变化对价格的影响。罗森茨韦格和沃尔平(Rosenzweig and Wolpin，2000)回顾建立在自然实验譬如双胞胎实验基础上的广泛文献。艾萨克森(Isacsson，1999)利用瑞典人数据，使用双胞胎方法研究了受教育回报。安格里斯特和拉维(Angrist and Lavy，1999)研究班级大小对测验的影响，利用受限于“迈蒙尼德斯规则”(Maimonides' Rule，有关内容将在25.6节简略评述)的学校数据，他们认为，班级的大小不应超过40人。该规则生成一个工具。

第二部分
核心方法

第二部分表述核心估计方法——最小二乘法、极大似然法和矩方法，以及与处于微观经济计量学中心地位的非线性回归模型有关的推断方法。内容还包括一些现代专题，譬如分位数回归、序贯估计、经验似然、半参数回归与非参数回归，以及基于自助法的统计推断。通常，讨论打算提供充足的背景和详细内容，以便使实践者可以阅读与领会重要经济计量学期刊上的论文，并且提供本书后续章节所需要的知识。我们假定，读者已熟悉线性回归分析。

有三章内容阐述基本估计理论。第 4 章以线性回归模型开始。然后，在导论水平上介绍分位数回归，它是对分布特性而不是条件均值进行建模。它提供了对工具变量很长的阐述处理，这是因果推断的主要方法。第 5 章阐述非线性模型的最广泛使用的估计方法，在对极大似然及非线性最小二乘法回归专门化之前，以 m 估计专题来开始。第 6 章提供广义矩方法的综合处理，这是一种极具一般性的估计框架，它可应用于单方程与多方程背景下的线性和非线性模型。本章强调工具变量估计的特殊情况。

然后，我们回到模型检验上。第 7 章涵盖既经典又包含自助法假设检验的方法，而第 8 章阐述相对现代的模型选择方法与设定分析。由于这些内容的重要性，密集计算自助法也是第三部分中第 11 章阐述的更详细内容的主题。本书的独具特色之处是，检验方法尽量只以这三章的统一方式阐述。然后，本书自始至终地在特定应用中来阐述上述程序。

第 9 章是独特的一章，它阐述可以把灵活结构放置在经济计量模型上的非参数和半参数的估计方法。

第 10 章阐述用于计算第 5 章和第 6 章所表述的非线性估计量的计算方法。如果估计量不能通过经济计量学软件包自动计算，或者在模型估计中遇到数值求解困难，那么实践者就要特别地利用本章内容。

4 线性模型

4.1 引　论

在微观经济计量学中，大量实证研究都使用线性回归及其各种变形。在进入本书重点内容即非线性回归之前，我们将提供关于横截面数据的单方程线性回归模型的某些重要结果的概览。有关线性回归模型的几种不同的估计量也将加以阐述。

尤其是普通最小二乘(OLS)估计特别受到人们喜爱。对于一般的微观经济计量横截面数据模型来说，模型误差项可能是异方差的。于是，就异方差误差而言，统计推断应该是稳健的，并且，通过使用加权最小二乘法而不是 OLS 来获得有效性提高是可行的。

OLS 估计量是对残差平方和求最小值。一种可选择的方法是对残差绝对值之和求最小值，从而得到最小绝对偏差估计量。此估计量连同分位数回归的推广，也将得到阐述。

对各种模型的错误设定会导致最小二乘估计量的非一致性。在这些情况下，对经济上关注的参数进行推断就需要更高等的方法，本书将对这些方法进行详细而深入的阐述。一种普遍使用的方法是工具变量回归。本章将对该重要方法提供一个介绍性研究，并且讨论弱工具的含义。

4.2 节提供回归定义，并阐述各种损失函数，从而引出回归函数的各种不同的估计量。4.3 节给出一个事例。一些重要的估计方法，尤其是普通最小二乘法、加权最小二乘法以及分位数回归，分别在 4.4 节、4.5 节以及 4.6 节加以阐述。对模型错误设定将在 4.7 节考察。4.8 节与 4.9 节均阐述工具变量回归。4.3 节～4.5 节、4.7 节以及 4.8 节将涵盖引论课程中的标准内容，而 4.2 节、4.6 节以及 4.9 节则引进更高等的内容。

4.2　回归与损失函数

在现代微观经济计量学中，“回归”(**regression**)这一术语意指，研究结果变量 y 与一系列回归元 $\mathbf{x}$ 之间关系的众多方法。因此，阐述某些重要回归形式是有益的。

为了解释方便，考察给定 $\mathbf{x}$ 时回归作为 y 的条件预测(**conditional prediction**)的目的。在实际应用中，回归模型还用于其他目的，尤其是因果关系的推断。尽管

这样，预测函数是对有用数据的一种概括，同时仍然是关注的内容。特别地，参见4.2.3节中关于线性预测和建立在线性因果均值上的因果推断之间的差异。

4.2.1 *损失函数*

设$\hat{y}$表示预测量[1]（**predictor**），把它定义成$\mathbf{x}$的函数。设$e \equiv y-\hat{y}$表示预测误差（**prediction error**），并设：

$$\mathrm{L}(e)=\mathrm{L}(y-\hat{y}) \tag{4.1}$$

表示与误差e有关的损失（**loss**）。如同在决策分析中，我们假定预测量构成某一决策的基础，而预测误差则会导致决策者的不利[2]（**disutility**），这由$\mathrm{L}(e)$所刻画，$\mathrm{L}(e)$的精确函数形式是由决策者来选择的。损失函数具有随$|e|$增大而递增的特性。一旦把$(y,\ \hat{y})$处理成随机的，对决策者损失函数的期望值求最小值，记为$\mathrm{E}[\mathrm{L}(e)]$。如果预测量依赖于$K$维向量$\mathbf{x}$，那么期望损失（**expected loss**）可表述成：

$$\mathrm{E}[\mathrm{L}((y-\hat{y})|\mathbf{x})] \tag{4.2}$$

对损失函数的选择，本质上应依赖于与预测误差有关的损失。在一些场合，譬如天气预报，也许存在关于选取一种损失函数而不是其他函数的一个可靠基础。

在经济计量学中，往往不存在显而易见的指南，习惯上是设定二次损失。于是，把式（4.1）专门化为$\mathrm{L}(e)=e^2$，同时通过式（4.2），最优预测量是对期望误差$\mathrm{E}[\mathrm{L}(e|\mathbf{x})]=\mathrm{E}[e^2|\mathbf{x}]$求最小值。由此可得，在这种情况下，最小均方预测误差准则常用于比较预测量。

4.2.2 *最优预测*

选择最优预测量（**optimal predictor**）的决策理论是通过最小化期望损失（**minimizing expected loss**）

$$\min_{\hat{y}} \mathrm{E}[\mathrm{L}((y-\hat{y})|\mathbf{x})]$$

构成的。因此，最优性的性质是与决策者的损失函数有关的。

表4.1给出损失函数的四个重要事例，以及相关的最优预测量函数。我们依次对每一种方法提供一个简要的介绍。详细分析已由曼斯基（Manski，1988a）给出。

表4.1　损失函数与对应的最优预测式

损失函数类型	定　义	最优预测式
平方误差损失	$\mathrm{L}(e)=e^2$	$\mathrm{E}[y\|\mathbf{x}]$
绝对误差损失	$\mathrm{L}(e)=\|e\|$	$\mathrm{med}[y\|\mathbf{x}]$
非对称绝对损失	$\mathrm{L}(e)=\begin{cases}(1-\alpha)\|e\|, & \text{如果 } e<0 \\ \alpha\|e\|, & \text{如果 } e\geqslant 0\end{cases}$	$q_\alpha[y\|\mathbf{x}]$
分步损失	$\mathrm{L}(e)=\begin{cases}0, & \text{如果 } e<0 \\ 1, & \text{如果 } e\geqslant 0\end{cases}$	$\mathrm{mod}[y\|\mathbf{x}]$

〔1〕又称为预测元或预测式。——译者注
〔2〕又称为无效性。——译者注

最有名的损失函数是平方误差损失(**squared error loss**,或称均方损失)函数。于是,y 的最优预测量是条件均值函数(**conditional mean function**)$E[y|\mathbf{x}]$。在绝大多数情况下,对 $E[y|\mathbf{x}]$不施加任何结构,而且可通过非参数回归(参见第 9 章)加以估计。在许多情况下,$E[y|\mathbf{x}]$的模型是设定好的,满足 $E[y|\mathbf{x}]=g(\mathbf{x}, \boldsymbol{\beta})$,其中,$g(\cdot)$表示已设定函数,而 $\boldsymbol{\beta}$ 表示需要估计的有限维参数向量。最优预测是$\hat{y}=g(\mathbf{x}, \hat{\boldsymbol{\beta}})$,其中,$\hat{\boldsymbol{\beta}}$ 表示求样本损失最小值的选择:

$$\sum_{i=1}^{n} \mathrm{L}(e_i) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - g(\mathbf{x}_i, \boldsymbol{\beta}))^2$$

损失函数是残差平方之和,所以可通过非线性最小二乘(参见 5.8 节)进行估计。如果把条件均值函数 $g(\cdot)$限定成关于 $\mathbf{x}$ 与 $\boldsymbol{\beta}$是线性的,因此 $E[y|\mathbf{x}]=\mathbf{x}'\boldsymbol{\beta}$,那么最优预测量为$\hat{y}=\mathbf{x}'\hat{\boldsymbol{\beta}}$,其中,$\hat{\boldsymbol{\beta}}$表示 4.4 节将要详述的普通最小二乘法估计量。

若损失函数是绝对误差损失(**absolute error loss**),则最优预测量是条件中位数(**conditional median**),记为 $\mathrm{med}[y|\mathbf{x}]$。如果条件中位数函数是线性的,因而 $\mathrm{med}[y|\mathbf{x}]=\mathbf{x}'\boldsymbol{\beta}$,那么最优预测量是$\hat{y}=\mathbf{x}'\hat{\boldsymbol{\beta}}$,其中,$\hat{\boldsymbol{\beta}}$ 表示求$\sum_i|y_i-\mathbf{x}_i'\boldsymbol{\beta}|$最小值的最小绝对离差估计量。4.6 节将阐述这个估计量。

平方误差损失函数是对称的,绝对误差损失函数也是对称的,不管预测误差的方向如何,就给定量值的预测误差而言,均可利用相同的惩罚(**penalty**)。相反,非对称绝对误差损失(**asymmetric absolute error loss**)对过度预测施加惩罚 $(1-\alpha)|e|$,而对过低预测施加不同的惩罚 $\alpha|e|$ 。非对称参数 α 是设定的。当 $\alpha=0.5$ 时,它位于区间(0, 1)之间且是对称的;然而,当 α 接近 0 或 1 时,就增大了非对称性。可以证明,最优预测量是条件分位数(**conditional quantile**),记为$q_\alpha[y|\mathbf{x}]$;一种特殊情况是,当 $\alpha=0.5$ 时为条件中位数。4.6 节将定义条件分位数,并且阐述分位数回归[凯恩克和巴西特(Koenker and Bassett, 1978)]。

表 4.1 给出的最后一个损失函数是阶跃损失(**step loss**),它把损失直接建立在预测误差的符号基础上,而不管其量值如何。最优预测值是条件众数,记为 $\mathrm{mod}[y|\mathbf{x}]$。这就提供了众数回归的动机[李(Lee, 1989)]。

极大似然估计并没有如此简单地进入本节的预测框架。然而,它能被给予以预测密度和最小化库尔贝克—利伯勒信息的形式对期望损失的解释。

所表述的结果蕴含着,经济计量学家对从数据(y, $\mathbf{x}$)中估计预测函数必须依据损失函数来选择预测函数感兴趣。运用流行的线性回归至少隐含地表明,决策者拥有一个二次损失函数,并认为条件均值函数是线性的。然而,一旦设定其他三个种损失函数之一,那么最优预测量也将建立在该种类型的基础上。在实际应用中,并没有明显的理由去偏爱哪个特定的损失函数。

回归经常用于对数据的概括归纳,而不是为了特定预测本身。于是,当可选择的估计量可以提供关于估计灵敏度的有用信息时,考察一系列估计量是有益的。曼斯基(Manski, 1988a, 1991)曾指出,二次误差损失函数与绝对误差损失函数都是凸的。如果 $y|\mathbf{x}$ 的条件分布是对称的,那么条件均值与中位数估计量两者均是一致的,而且可以认为,它们是相当接近的。此外,如果人们要避免关于 $y|\mathbf{x}$ 的分布假设,那么可供选择的估计量方面的差异提供了一种认识数据分布的途径。

4.2.3 线性预测

在平方误差损失下，最优预测值是条件均值 $\mathrm{E}[y|\mathbf{x}]$。如果这个条件均值关于 $\mathbf{x}$ 是线性的，因而 $\mathrm{E}[y|\mathbf{x}]=\mathbf{x}'\boldsymbol{\beta}$，那么参数 $\boldsymbol{\beta}$ 具有结构解释或者因果解释，通过 OLS 得到的 $\boldsymbol{\beta}$ 的一致估计，蕴含着 $\mathrm{E}[y|\mathbf{x}]=\mathbf{x}'\boldsymbol{\beta}$ 的一致估计。这允许关于回归元的变动对条件均值的效应，进行有意义的政策分析。

然而，若条件均值关于 $\mathbf{x}$ 是非线性的，因而 $\mathrm{E}[y|\mathbf{x}]\neq\mathbf{x}'\boldsymbol{\beta}$，则 OLS 的结构解释就会消失。然而，在平方误差损失下，把 $\boldsymbol{\beta}$ 解释成最优线性预测量，仍然是可行的。一旦求期望损失 $\mathrm{E}[(y-\mathbf{x}'\boldsymbol{\beta})^2]$ 关于 $\boldsymbol{\beta}$ 的导数，可得到一阶条件 $-2\mathrm{E}[\mathbf{x}(y-\mathbf{x}'\boldsymbol{\beta})]=\mathbf{0}$，因此，最优线性预测量是 $\boldsymbol{\beta}=(\mathrm{E}[\mathbf{xx}'])^{-1}\mathrm{E}[\mathbf{x}y]$，与样本的 OLS 估计量类似。

通常，我们专门研究带有截距项的模型。在对记号进行变动后，我们把 $\mathbf{x}$ 定义成排除截距的回归元，并用 $\alpha+\mathbf{x}'\boldsymbol{\gamma}$ 代替 $\mathbf{x}'\boldsymbol{\beta}$。关于 α 与 $\boldsymbol{\gamma}$ 的一阶条件是 $-2\mathrm{E}[u]=0$ 与 $-2\mathrm{E}[\mathbf{x}u]=\mathbf{0}$，其中，$u=y-(\alpha+\mathbf{x}'\boldsymbol{\gamma})$。这些条件蕴含着 $\mathrm{E}[u]=0$ 且 $\mathrm{Cov}[\mathbf{x},u]=\mathbf{0}$。经过求解，得出：

$$\begin{aligned}\boldsymbol{\gamma}&=(\mathrm{V}[\mathbf{x}])^{-1}\mathrm{Cov}[\mathbf{x},y]\\ \alpha&=\mathrm{E}[y]-\mathrm{E}[\mathbf{x}']\boldsymbol{\gamma}\end{aligned}\tag{4.3}$$

例如，参见戈德伯格(Goldberger，1991，第 52 页)。

从式(4.3)推导得知，应该很明显，对于数据$(y,\ \mathbf{x})$，我们总能把线性回归模型写成：

$$y=\alpha+\mathbf{x}'\boldsymbol{\gamma}+u\tag{4.4}$$

其中，参数 α 与 $\boldsymbol{\gamma}$ 都已在式(4.3)中定义，而误差项 u 满足 $\mathrm{E}[u]=0$ 和 $\mathrm{Cov}[\mathbf{x},u]=\mathbf{0}$。

因此，在平方误差损失下，总是可以给出线性回归模型作为最优线性预测(**best linear prediction**)或线性投影的非结构或简化式解释。然而，因为条件均值关于 $\mathbf{x}$ 是线性的，因此 $\mathrm{E}[y|\mathbf{x}]=\alpha+\mathbf{x}'\boldsymbol{\gamma}$，要求假设 $\mathrm{E}[u|\mathbf{x}]=0$，并且有 $\mathrm{E}[u]=0$ 和 $\mathrm{Cov}[\mathbf{x},u]=\mathbf{0}$。

这种差异具有实践上的重要意义。例如，若 $\mathrm{E}[u|\mathbf{x}]=0$，因而 $\mathrm{E}[y|\mathbf{x}]=\alpha+\mathbf{x}'\boldsymbol{\gamma}$，则最小二乘(LS)估计量 $\hat{\boldsymbol{\gamma}}$ 的概率极限是 $\boldsymbol{\gamma}$，而不管 LS 估计量是加权的还是未加权的，也不管样本是通过简单随机抽样还是通过外生分层抽样而得到。然而，如果 $\mathrm{E}[y|\mathbf{x}]\neq\alpha+\mathbf{x}'\boldsymbol{\gamma}$，那么这些不同的 LS 估计量可能具有不同的概率极限。这种事例将在 24.3 节进一步讨论。

OLS 的结构性解释，需要在给定回归元时误差项的条件均值等于 0。

4.3 例子：受教育回报

在劳动经济学中，重要的线性回归应用涉及测算教育对工资或薪水的影响。一个典型的受教育回报(**returns to schooling**)模型设定：

$$\ln w_i=\alpha s_i+\mathbf{x}_{2i}'\boldsymbol{\beta}+u_i,\quad i=1,\cdots,N\tag{4.5}$$

其中，w 表示小时工资或年薪，s 表示所完成的受教育年数，$\mathbf{x}_2$ 表示控制变量，譬如工作经验、性别或家庭背景等。下标 i 代表样本中的第 i 个人。由于因变量是工资对数，所以模型是一个对数线性模型，系数 α 测算了与多受一年教育相联系的薪水的比例变化。

这个模型中经常使用的估计方法是普通最小二乘法。在实际应用中，对 $\ln w$ 变换确保了误差大致上是同方差的，但是，最好仍然如同 4.4 节所述的那样，去获得异方差一致标准误差。如果关注内容为分布问题，比如下四分位的特性，那么还可通过分位数回归进行估计(参见 4.6 节)。

回归(4.5)可立即以描述方式得到应用。例如，如果$\hat{\alpha}=0.10$，一旦 $\mathbf{x}_2$ 中包括的全部因素得到控制，那么受一年教育就会有 10%的薪水变化。如同本例一样，重要的是，添加最后一项限定时，由于 $\mathbf{x}_2$ 包括另外可能影响薪水的控制因素，譬如收入影响，估计量$\hat{\alpha}$ 通常变得较小。

政策上关注的内容在于，确定受教育方面的外生变化对薪水的影响。然而，受教育并不是随机指派的，相反，它依赖于个体者所做出的选择。人力资本理论认为，受教育是个体者自身投资，而 α 被解释成对人力资本回报的测算。于是，回归(4.5)是单个内生变量 $\ln w$ 对另一个变量 s 的回归，因而它不能测算出 s 外生变动的因果影响。此处的条件均值函数并不具有因果意义，因为它是以受教育作为内生因素为条件的。实际上，除非我们能证明，s 本身是一些变量的函数，而在这些变量中，至少有一个变量可以独立于 u 而变化，否则，把 α 看成因果参数，它所蕴含的内容并不清楚。

这种带有个体可观测数据的内生回归元，遍及微观经济计量学分析之中。由 4.4 节给出的线性回归模型的标准假设为：回归元均是外生的。内生回归元的后果将在 4.7 节考察。控制外生回归元的一种方法是工具变量，这将在 4.8 节详述。最近，安格里斯特和克鲁格(Angrist and Krueger，1999)给出对这类工资—受教育事例中控制内生变量方法的广泛评述。这些方法已在 2.8 节给出一个概述，并且将贯穿于全书中。

4.4 普通最小二乘法

在线性回归模型中，最简单的回归事例是 OLS 估计量。

首先定义模型与估计量，然后给出对 OLS 估计量渐近分布的详细阐述。此处的阐述假定，前面的表述是一种更具引导性的处理。这里做出的模型假设，允许随机回归元和异方差误差，同时建议数据要通过外生分层抽样来得到。

关于怎样获得 OLS 估计量的异方差性稳健标准误差的关键结果，将在 4.4.5 节给出。

4.4.1 *线性回归模型*

在标准截面数据回归模型中，具有一个纯量因变量与几个回归元的 N 个观测值，该数据被设定成($\mathbf{y}$, $\mathbf{X}$)，其中，$\mathbf{y}$ 表示因变量的观测值，而 $\mathbf{X}$ 表示解释变量的

矩阵。

具有可加误差的一般回归模型可用向量形式写成：

$$\mathbf{y}=\mathrm{E}[\mathbf{y}|\mathbf{X}]+\mathbf{u} \tag{4.6}$$

其中，$\mathbf{E}[\mathbf{y}|\mathbf{X}]$表示在给定 $\mathbf{X}$ 时 $\mathbf{y}$ 的条件期望，而 $\mathbf{u}$ 表示不可观测的随机误差或分布向量。此方程式右边把 $\mathbf{y}$ 分解为两种成分：一种是确定性已知的回归元，另一种起因于随机变动或噪声。我们把 $\mathrm{E}[\mathbf{y}|\mathbf{X}]$看成条件预测函数，它会产生平均值，或者更正式地，为给定 $\mathbf{X}$ 时 $\mathbf{y}$ 的期望值。

当 $\mathrm{E}[\mathbf{y}|\mathbf{X}]$被设定成 $\mathbf{X}$ 的线性函数时，就得到一个线性回归模型（**linear regression model**）。关于此模型的记号，已在 1.6 节中详细介绍过。以向量形式表示第 i 个观测为：

$$y_i=\mathbf{x}_i'\boldsymbol{\beta}+u_i \tag{4.7}$$

其中，$\mathbf{x}_i$ 表示 $K\times1$ 维回归元向量（**regressor vector**），$\boldsymbol{\beta}$表示 $K\times1$ 维参数向量（**parameter vector**）。有时候，比较简单的是省略下标 i，而把典型观测数据的模型写成 $y=\mathbf{x}'\boldsymbol{\beta}+u$。在矩阵中 N 个观测值排列成行，得出：

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{u} \tag{4.8}$$

其中，$\mathbf{y}$ 表示 $N\times1$ 维因变量向量（**dependent variable vector**），$\mathbf{X}$ 表示 $N\times K$ 阶回归元矩阵（**regressor matrix**），$\mathbf{u}$ 表示 $N\times1$ 维误差向量（**parameter vector**）。

对于线性回归模型来说，方程(4.7)和(4.8)是等价的，并且两者可交换使用。后者更为简洁，经常是最方便的表述形式。

在这种背景下，y 称为因变量（**dependent variable**）或内生变量（**endogenous variable**），我们希望用 $\mathbf{x}$ 与 u 的变化研究 y 的变动；u 称为误差项（**error term**）或干扰项（**disturbance term**）；$\mathbf{x}$ 称为回归元（**regressors**）、预测量（**predictors**）或协变量（**covariates**）。如果 4.4.6 节中的假设 4 成立，那么 $\mathbf{x}$ 的所有分量都是外生变量（**exogenous variable**）或自变量（**independent variable**）。

4.4.2 *OLS 估计量*

OLS 估计量被定义为对误差平方和：

$$\sum_{i=1}^{n}u_i^2=\mathbf{u}'\mathbf{u}=(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \tag{4.9}$$

求最小值的估计量。令式(4.9)关于 $\boldsymbol{\beta}$的导数等于 0，并且求解 $\boldsymbol{\beta}$，得到 OLS 估计量：

$$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{4.10}$$

更一般的结果参见习题 4.5，它假定矩阵 $\mathbf{X}'\mathbf{X}$ 的逆存在。若 $\mathbf{X}'\mathbf{X}$ 是非满秩的，则其逆矩阵可用广义逆来代替。于是，如果使用平方误差损失，那么通过 OLS 估计，仍然可得出在给定 $\mathbf{x}$ 时 y 的最优线性预测，只是 $\mathbf{x}$ 的各种不同线性组合将会产生这个最优预测量。

4.4.3 识别

倘若 $\mathbf{X}'\mathbf{X}$ 是非奇异的，OLS 估计量就总是能够计算出来。更加令人关注的问题是，$\hat{\beta}_{OLS}$会告诉我们有关数据的什么内容？

为了使条件均值 $E[\mathbf{y}|\mathbf{X}]$的识别成为可能（参见 2.5 节），我们关注 OLS 估计量的能力。对于线性模型，参数 β 可识别的，如果：

1. $E[\mathbf{y}|\mathbf{X}]=\mathbf{X}\beta$；
2. $\mathbf{X}\beta^{(1)}=\mathbf{X}\beta^{(2)}$，当且仅当 $\beta^{(1)}=\beta^{(2)}$。

第 1 个条件是，条件均值被正确设定，它确保 β 成为内在的关注内容；第 2 个假设蕴含着 $\mathbf{X}'\mathbf{X}$ 是非奇异的，这与计算唯一的 OLS 估计值(4.10)所需要的条件相同。

4.4.4 OLS 估计量的分布

我们集中考虑 OLS 估计量的渐近性质。首先建立一致性，然后通过对 OLS 估计量重新标度获得其极限分布。随后，统计推断要求对估计量方差矩阵的一致估计。这一分析广泛利用渐近理论，附录 A 将概述渐近理论。

一致性

估计量的性质依赖于真实生成数据的过程，也就是数据生成过程(**data generating process, dgp**)。假定 dgp 是 $\mathbf{y}=\mathbf{X}\beta+\mathbf{u}$，因而，模型(4.8)是正确设定的。在一些地方，尤其是在第 5 章、第 6 章和附录 A，把下标 0 添加到 β 上，因此，dgp 变成 $\mathbf{y}=\mathbf{X}\beta_0+\mathbf{u}$。更多的讨论参见 5.2.3 节。

于是：

$$\begin{aligned}\hat{\beta}_{OLS}&=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\\&=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta+\mathbf{u})\\&=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\end{aligned}$$

从而，OLS 估计量可表示成：

$$\hat{\beta}_{OLS}=\beta+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \tag{4.11}$$

为了证明一致性，我们重新把式(4.11)写成：

$$\hat{\beta}_{OLS}=\beta+(N^{-1}\mathbf{X}'\mathbf{X})^{-1}N^{-1}\mathbf{X}'\mathbf{u} \tag{4.12}$$

对等号右边重新正规化的原因是，如果 $\mathbf{x}_i$ 满足允许对 $\mathbf{x}_i\mathbf{x}_i'$ 应用大数定理的假设(详细内容参见 4.4.8 节)，$N^{-1}\mathbf{X}'\mathbf{X}=N^{-1}\sum_i\mathbf{x}_i\mathbf{x}_i'$ 依概率收敛到有限非零矩阵的平均值。进而利用斯拉茨基(Slutsky)定理(定理 A.3)：

$$\text{plim}\ \hat{\beta}_{OLS}=\beta+(\text{plim}\ N^{-1}\mathbf{X}'\mathbf{X})^{-1}(\text{plim}\ N^{-1}\mathbf{X}'\mathbf{u})$$

OLS 估计量关于 β是一致的(也就是说，$\text{plim}\ \hat{\beta}_{OLS}=\beta$)，如果：

$$\text{plim}\ N^{-1}\mathbf{X}'\mathbf{u}=\mathbf{0} \tag{4.13}$$

若大数定律应用于平均数 $N^{-1}\mathbf{X}'\mathbf{u}=N^{-1}\sum_i\mathbf{x}_iu_i$ 上，则使式(4.13)成立的必要条件是 $E[\mathbf{x}_iu_i]=\mathbf{0}$。

极限分布

给定一致性，$\hat{\beta}_{OLS}$的极限分布退化，并且所有质量位于β处。为了获得该极限分布，我们用$\sqrt{N}$乘$\hat{\beta}_{OLS}$，因为这种重新标度会导致随机变量在标准横截面数据的假设下渐近地具有非零且有限的方差。于是，式(4.11)变为：

$$\sqrt{N}(\hat{\beta}_{OLS}-\beta)=(N^{-1}\mathbf{X}'\mathbf{X})^{-1}N^{-1/2}\mathbf{X}'\mathbf{u} \tag{4.14}$$

一致性的证明假定，plim $N^{-1}\mathbf{X}'\mathbf{X}$存在，并且是有限且非零的。我们假定中心极限定理可应用于$N^{-1/2}\mathbf{X}'\mathbf{u}$，得到多变量正态极限分布，具有有限、非奇异的协方差矩阵。若对极限正态分布应用乘积法则(定理A.17)，则蕴含着式(4.14)右边的乘积具有极限正态分布。详细内容由4.4.8节给出。

这就得出以下命题，它允许回归元是随机的，同时没有把模型误差限制成同方差的。

命题4.1(OLS估计量分布) 做出下述假设：

(i) dgp是模型(4.8)，即$\mathbf{y}=\mathbf{X}\beta+\mathbf{u}$；

(ii) 对于不同的i，数据是独立的，满足$\mathrm{E}[\mathbf{u}|\mathbf{X}]=\mathbf{0}$，$\mathrm{E}[\mathbf{uu}'|\mathbf{X}]=\mathbf{\Omega}=\mathrm{Diag}[\sigma_i^2]$；

(iii) 矩阵$\mathbf{X}$满秩的，因而$\mathbf{X}\beta^{(1)}=\mathbf{X}\beta^{(2)}$，当且仅当$\beta^{(1)}=\beta^{(2)}$；

(iv) $K\times K$阶矩阵

$$\mathbf{M}_{\mathbf{XX}}=\mathrm{plim}\,N^{-1}\mathbf{X}'\mathbf{X}=\mathrm{plim}\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'=\lim\frac{1}{N}\sum_{i=1}^{N}\mathrm{E}[\mathbf{x}_i\mathbf{x}_i'] \tag{4.15}$$

存在，而且是有限非奇异的。

(v) $K\times 1$维向量$N^{-1/2}\mathbf{X}'\mathbf{u}=N^{-1/2}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{u}_i\xrightarrow{d}\mathcal{N}[\mathbf{0},\ \mathbf{M}_{\mathbf{X\Omega X}}]$

其中：

$$\mathbf{M}_{\mathbf{X\Omega X}}=\mathrm{plim}\,N^{-1}\mathbf{X}'\mathbf{uu}'\mathbf{X}=\mathrm{plim}\frac{1}{N}\sum_{i=1}^{N}u_i^2\mathbf{x}_i\mathbf{x}_i'=\lim\frac{1}{N}\sum_{i=1}^{N}\mathrm{E}[u_i^2\mathbf{x}_i\mathbf{x}_i'] \tag{4.16}$$

则由式(4.10)定义的OLS估计量$\hat{\beta}_{OLS}$是关于β一致的，且：

$$\sqrt{N}(\hat{\beta}_{OLS}-\beta)\xrightarrow{d}\mathcal{N}[\mathbf{0},\ \mathbf{M}_{\mathbf{XX}}^{-1}\mathbf{M}_{\mathbf{X\Omega X}}\mathbf{M}_{\mathbf{XX}}^{-1}] \tag{4.17}$$

假设(i)用于获得式(4.11)。假设(ii)确保$\mathrm{E}[\mathbf{y}|\mathbf{X}]=\mathbf{X}\beta$，同时使得方差$\sigma_i^2$具有异方差误差，这比限制$\mathbf{\Omega}=\sigma^2\mathbf{I}$的同方差不相关的误差更具一般性。假设(iii)排除了回归元之间的完全共线性。假设(iv)导致用N^{-1}对式(4.12)与式(4.14)中的$\mathbf{X}'\mathbf{X}$重新标度。注意，利用大数定理，有plim=lim E(参见附录4.3节)。

一致性的根本条件是式(4.13)。我们不是直接假定这一条件，而是使用更强的假设(v)，它是获得式(4.17)所必需的。倘若$N^{-1/2}\mathbf{X}'\mathbf{u}$具有极限分布，其均值为零且方差有限，乘以$N^{-1/2}$得出依概率收敛到0的随机变量，因此，正如人们所期望的，式(4.13)成立。关于u_i与$\mathbf{x}_i$的更原始假设，确保(iv)与(v)得到满足，这些由4.4.6节给出，而正式证明则放在4.4.8节。

渐近分布

命题 4.1 给出$\sqrt{N}(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}-\boldsymbol{\beta})$的极限分布(**limit distribution**),即$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$的重新标度形式。许多实践者更愿意看到用$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$分布直接写成的渐近结果,在此情况下,这种分布称为渐近分布(**asymptotic distribution**)。这种渐近分布应用于大样本(**large sample**),意味着样本足够大到使极限分布得到良好近似,但是没有大到使$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\xrightarrow{p}\boldsymbol{\beta}$,进而它的渐近分布变成退化的。这方面讨论放在附录 A.6.4 中。

渐近分布由式(4.17)除以$\sqrt{N}$,并且加上$\boldsymbol{\beta}$获得。这就得出渐近分布(**asymptotic distribution**):

$$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\overset{a}{\sim}\mathcal{N}[\boldsymbol{\beta},N^{-1}\mathbf{M}_{\mathbf{xx}}^{-1}\mathbf{M}_{\mathbf{x\Omega x}}\mathbf{M}_{\mathbf{xx}}^{-1}] \tag{4.18}$$

其中,符号$\overset{a}{\sim}$表示“在渐近形式上分布为”。式(4.18)中的方差矩阵称为$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$的渐近方差矩阵(**asymptotic variance matrix**),并用$\mathrm{V}[\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}]$表示。更简单的记号是在对$\mathbf{M}_{\mathbf{xx}}$与$\mathbf{M}_{\mathbf{x\Omega x}}$的定义中省略极限与期望,从而渐近分布记为:

$$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\overset{a}{\sim}\mathcal{N}[\boldsymbol{\beta},(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \tag{4.19}$$

而$\mathrm{V}[\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}]$定义成式(4.19)中的方差矩阵。

在后面一些章节中,我们既使用式(4.18)又使用式(4.19)来表述渐近分布。使用它们是为了便于表述。统计推断的正式渐近结果建立在极限分布而不是渐近分布的基础上。

为了具体推导,式(4.17)与式(4.18)中的矩阵$\mathbf{M}_{\mathbf{xx}}$与$\mathbf{M}_{\mathbf{x\Omega x}}$都要用一致估计$\hat{\mathbf{M}}_{\mathbf{xx}}$与$\hat{\mathbf{M}}_{\mathbf{x\Omega x}}$代替。于是,$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$的估计渐近方差矩阵(**estimate asymptotic variance matrix**)是:

$$\hat{\mathrm{V}}[\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}]=N^{-1}\hat{\mathbf{M}}_{\mathbf{xx}}^{-1}\hat{\mathbf{M}}_{\mathbf{x\Omega x}}\hat{\mathbf{M}}_{\mathbf{xx}}^{-1} \tag{4.20}$$

这个估计量称为三明治估计(**sandwich estimate**),$\hat{\mathbf{M}}_{\mathbf{x\Omega x}}$夹在$\hat{\mathbf{M}}_{\mathbf{xx}}^{-1}$与$\hat{\mathbf{M}}_{\mathbf{xx}}^{-1}$中间。

4.4.5 *OLS* 的异方差稳健标准误差

在式(4.20)中,对$\hat{\mathbf{M}}_{\mathbf{xx}}$的明显选择应该是$N^{-1}\mathbf{X}'\mathbf{X}$。对式(4.16)定义的$\mathbf{M}_{\mathbf{x\Omega x}}$进行估计,依赖于对误差项所做出的假设。

在微观经济计量应用中,模型误差经常是条件异方差的,即$\mathrm{V}[u_i|x_i]=\mathrm{E}[u_i^2|x_i]=\sigma_i^2$,其中,$\sigma_i^2$随$i$而变化。怀特(White, 1980a)建议利用$\hat{\mathbf{M}}_{\mathbf{x\Omega x}}=N^{-1}\sum_i\hat{u}_i^2\mathbf{x}_i\mathbf{x}_i'$,这种估计需要由 4.4.8 节给出的额外假设。

一旦把估计值$\hat{\mathbf{M}}_{\mathbf{xx}}$与$\hat{\mathbf{M}}_{\mathbf{x\Omega x}}$加以结合,并进行简化,就得到估计渐近方差矩阵的估计值:

$$\begin{aligned}\hat{\mathrm{V}}[\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}]&=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\boldsymbol{\Omega}}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\\&=\Big(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\Big)^{-1}\sum_{i=1}^{N}\hat{u}_i^2\mathbf{x}_i\mathbf{x}_i'\Big(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\Big)^{-1}\end{aligned} \tag{4.21}$$

其中,$\hat{\boldsymbol{\Omega}}=\mathrm{Diag}[\hat{u}_i^2]$,而$\hat{u}_i=y_i-\mathbf{x}_i'\hat{\boldsymbol{\beta}}$表示 OLS 残差。归功于怀特(White, 1980a)的

这个估计值，称为 OLS 估计量的渐近方差矩阵的异方差一致(**heteroskedasticity-consistent**)估计值，它所产生的标准误差称为异方差稳健标准误差(**heteroskedasticity-robust standard error**)，或者，更简单地称为稳健标准误差(**robust standard error**)。即使 $\hat{u}_i^2$ 对关于σ_i^2 不是一致的，也提供了 $\mathrm{V}[\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}]$的一致估计。

在引论中，误差被限制为同方差的(**homoskedastic**)。于是，$\boldsymbol{\Omega}=\boldsymbol{\sigma}^2\mathbf{I}$，因而 $\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}=\sigma^2\mathbf{X}'\mathbf{X}$，从而 $\mathbf{M}_{\mathbf{x}\Omega\mathbf{x}}=\sigma^2\mathbf{M}_{\mathbf{xx}}$。式(4.17)中的极限分布方差矩阵简化成 $\sigma^2\mathbf{M}_{\mathbf{XX}}^{-1}$，并且许多计算机软件包有时均使用所谓的默认 OLS 方差估计值：

$$\widetilde{\mathrm{V}}[\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}]=s^2(\mathbf{X}'\mathbf{X})^{-1} \tag{4.22}$$

其中，$s^2=(N-K)^{-1}\sum_i\hat{u}_i^2$。

建立在式(4.22)而不是式(4.21)基础上的推断是无效的，除非误差是同方差且不相关的。通常，当误差是异方差时，横截面数据经常出现此种情况，错误使用式(4.22)能引起高估真实标准误差或低估真实标准误差。

在实际应用中，利用$(N-K)$而不是用 N 去除，就可计算 $\hat{\mathbf{M}}_{\mathbf{x}\Omega\mathbf{x}}$，这类似于同方差情况下用 s^2 去除而得到一致估计值。那么，式(4.21)中的 $\hat{\mathrm{V}}[\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}]$要用 $N/(N-K)$去乘。就异方差误差而言，对此自由度的调整并不存在理论上的基础，但是，一些模拟研究却提供了支持[参见麦金农和怀特(Mackinnon and White, 1985)、朗和欧文(Long and Ervin, 2000)]。

在任何可能的情况下，微观经济计量学分析都要利用稳健标准误差。这种误差对异方差性而言是稳健的。防范其他错误设定同样是有把握的。特别地，当数据被聚集时，标准误差应是稳健的；参见 21.2.3 节与 24.5 节。

4.4.6 *截面数据回归的假设*

命题 4.1 是极具一般性的定理，它依赖于关于 $N^{-1}\mathbf{X}'\mathbf{X}$ 和 $N^{-1/2}\mathbf{X}'\mathbf{u}$ 的假设。在实际应用中，这些假设可通过对 $\mathbf{x}_i\mathbf{x}_i'$与 $\mathbf{x}_iu_i$ 的平均值应用大数定理以及中心极限定理得到验证。这些反过来需要观测值 $\mathbf{x}_i$ 与误差 u_i 是如何生成的假设，以及随后的式(4.7)中所定义的 y_i 是如何生成的。这些假设汇总起来，称为数据生成过程假设。

一个简单的教学例子由习题 4.4 给出。

我们在这一阶段的目标是做出适合于许多横截面数据应用背景的假设。一些假设曾经由怀特(White, 1980a)做出，还有违背引论课程中的那三个重要假设。首先，回归元可以是随机的(假设 1 与假设 3)，因此，做出关于误差项的假设是以回归元为条件的。其次，误差的条件方差对于不同观测值来说，可以是变化的(假设 5)。最后，误差不再被限制成正态分布的。

这些假设是：

1. 数据 $(y_i, \mathbf{x}_i)$对于不同 i 来说是独立的，且不是同分布的(inid)。

2. 模型是正确设定的，因此，有：

$$y_i=\mathbf{x}_i'\boldsymbol{\beta}_i+u_i$$

3. 回归元向量 $\mathbf{x}_i$ 可能是随机的，具有有限二阶矩。另外，对于某个 $\delta>0$ 以及所有的 j、$k=1,\cdots,K$，有 $\mathrm{E}[|x_{ij}x_{ik}|^{1+\delta}]\leqslant\infty$。同时，式(4.15)定义的矩阵 $\mathbf{M_{xx}}$ 存在，它是一个有限的正定矩阵，秩为 K。在分析样本时，$\mathbf{X}$ 的秩也为 K。

4. 误差具有零均值，并以回归元为条件：

$$\mathrm{E}[u_i|\mathbf{x}_i]=0$$

5. 误差以回归元为条件，是异方差的，满足：

$$\sigma_i^2=\mathrm{E}[u_i^2|\mathbf{x}_i] \tag{4.23}$$
$$\mathbf{\Omega}=\mathrm{E}[\mathbf{uu}'|\mathbf{X}]=\mathrm{Diag}[\sigma_i^2]$$

其中，$\mathbf{\Omega}$ 表示 $N\times N$ 阶有限的正定矩阵。同时，对于某个 $\delta>0$，有 $\mathrm{E}[|u_i^2|^{1+\delta}]<\infty$。

6. 由式(4.16)定义的矩阵 $\mathbf{M_{x\Omega x}}$ 存在，它是秩为 K 的有限正定矩阵，对于不同 i，$\mathbf{M_{x\Omega x}}=\mathrm{plim}\,N^{-1}\sum_i u_i^2\mathbf{x}_i\mathbf{x}_i'$ 具有独立性。同样地，对于某个 $\delta>0$ 以及所有 j、$k=1,\cdots,K$，有 $\mathrm{E}[|u_i^2x_{ij}x_{ik}|^{1+\delta}]<\infty$。

4.4.7 *假设评注*

为了完整起见，我们在下一节证明这些重要结果之前，对每一个假设提供详细讨论。

分层随机抽样

假设 1 常常是针对横截面数据以隐性方式提出的。在这里，我们以明确方式提出。它把数据 $(y_i,\ \mathbf{x}_i)$ 限制为对于不同 i 是独立的，但是允许分布随着 i 不同而不同。许多微观经济计量数据集合均来自分层随机抽样(**stratified random sampling**，参见 3.2 节)。于是，总体被分成一些层，然后从每一层内做随机抽取，但是，某些层作为 inid 的而非 iid 的抽样 $(y_i,\ \mathbf{x}_i)$，结果被过度抽取。相反，如果这些数据来自简单随机抽样(**simple random sampling**)，那么数据 $(y_i,\ \mathbf{x}_i)$ 就是 iid 的，这作为 inid 的特例是一种较强的假设。许多引论课程均假定，回归元在重复抽样中是固定的(**fixed in repeated samples**)。于是，数据 $(y_i,\ \mathbf{x}_i)$ 是 inid 的，因为唯一的 y_i 是随机的，它依赖于 $\mathbf{x}_i$ 的值。固定回归元假设极少适用于微观经济计量数据，此种数据通常是可观测的数据。相反，它可用于实验数据，$\mathbf{x}$ 表示处理水平。

这些关于 $(y_i,\ \mathbf{x}_i)$ 分布的各种不同假设，会影响到用大数定律和中心极限定理获得 OLS 估计量的渐近性质。注意到，即使 $(y_i,\ \mathbf{x}_i)$ 是 iid 的，给定 $\mathbf{x}_i$ 时，y_i 不是 iid 的，例如，$\mathrm{E}[y_i|\mathbf{x}_i]=\mathbf{x}_i'\boldsymbol{\beta}$ 随 $\mathbf{x}_i$ 而变化。

假设 1 排除了大多数时间序列数据，因为这些数据对于不同观测值是相关的。如果抽样方案包括聚集观测值，那么会违背假设 1。在这种情况下，倘若假设 2～假设 4 成立，OLS 估计量仍然是一致的，但是，通常它具有不同于本章所述的方差矩阵。

正确设定模型

假设 2 看起来非常明显，因为它在推导 OLS 估计量时作为根本性因素。然而，由于 $\hat{\boldsymbol{\beta}}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ 是关于 $\mathbf{y}$ 的函数，从而它的性质依赖于 $\mathbf{y}$。

如果假设 2 成立，那么假定回归模型关于 $\mathbf{x}$ 是线性的而不是非线性的，同时，

回归中没有省略变量(**omitted variables**),并且假定回归元不存在测量误差(**measurement error**),因为用于计算$\hat{\beta}$的回归元 $\mathbf{x}$ 与 dgp 中的回归元是一样的。再者,参数 β 对不同个体而言是相同的,这就排除了随机参数模型。

如果假设 2 得不到满足,那么 OLS 只可以被解释成一个最优线性预测量,参见 4.2.3 节。

随机回归元

假设 3 允许回归元可以为随机回归元(**stochastic regressor**),当使用调查数据而不是实验数据时,经常是这种情况。可以假定,在极限形式上,样本二阶矩矩阵是常数且为非奇异的。

若回归元是 iid 的,正如在简单随机抽样下所做出的假设,$\mathbf{M_{xx}}=\mathrm{E}[\mathbf{xx}']$,并且假设 3 被简化为二阶矩存在的假设。如果回归元是随机的且 inid 的,如同在分层随机抽样的情况下,我们就需要更强的假设 3,这个假设允许应用马尔可夫 LLN 来获得 $\mathrm{plim}\, N^{-1}\mathbf{X}'\mathbf{X}$。如果回归元在重复抽样中是固定的,即引论课程中做出稍欠满意的普遍假设,那么 $\mathbf{M_{xx}}=\lim N^{-1}\mathbf{X}'\mathbf{X}$,同时假设 3 变成这种极限存在的假设。

弱外生回归元

假设 4 的零条件均值误差是至关重要的,因为一旦它与假设 2 结合起来,就蕴含着 $\mathrm{E}[\mathbf{y}|\mathbf{X}]=\mathbf{X}\beta$,因此,条件均值实际上是 $\mathbf{X}\beta$。

假设 $\mathrm{E}[u|\mathbf{x}]=0$ 蕴含着 $\mathrm{Cov}[\mathbf{x},u]=\mathbf{0}$,因而误差与回归元是不相关的。接下来,由期望迭代定理可得出,$\mathrm{Cov}[\mathbf{x},u]=\mathrm{E}[\mathbf{x}u]-\mathrm{E}[\mathbf{x}]\mathrm{E}[u]$且 $\mathrm{E}[u|\mathbf{x}]=0$,蕴含着 $\mathrm{E}[\mathbf{x}u]=0$ 以及 $\mathrm{E}[u]=0$。比较弱的假设 $\mathrm{Cov}[\mathbf{x},u]=\mathbf{0}$ 就可以满足 OLS 一致性,然而,无偏的 OLS 则需要比较强的假设 $\mathrm{E}[u|\mathbf{x}]=0$ 。

假设 4 的经济意义是,误差项表述了所有被假定成与 $\mathbf{X}$ 不相关的外生因素,并且一般来说它们对 y 具有零影响。这是一个重要假设,在 2.3 节称为弱外生假设。在本质上,这意味着关于 $\mathbf{X}$ 变量的数据生成过程知识,对 β的估计并没有贡献什么有用的信息。当假设不能被满足时,K 回归元中的至少一个被称为与 y 是联合相关的(**jointly dependent**),或简称为内生的(**endogenous**)。回归元与误差相关的一般性术语是内生性(**endogeneity**)或内生回归元(**endogenous regressor**),其中,术语"内生"意味着由系统内的因素引起。正如我们在 4.7 节证明的,对弱外生性的违背会导致非一致估计量。存在许多方法违背弱外生性,但是,最普遍的一种方法是,$\mathbf{x}$ 中的变量是选择变量或决策变量,该变量在较大模型中与 y 相关。一旦忽视这些其他的关系,同时对 $\mathbf{x}_i$ 进行研究,就好像 $\mathbf{x}_i$ 被随机分配给观测值 i,从而与 u_i 不相关,这样做将得到非一般的结果。内生抽样(**endogenous sample**)已经由假设 4 排除了。然后,如果数据是由分层随机抽样收集起来的,那么它必定是外生分层抽样(**exogenous stratified**)。

条件异方差误差

假定独立回归误差(**independent regression error**)与回归元是不相关的,这是假设 1、假设 2 以及假设 4 的结果。引论课程通常关注把误差限制成同方差的,满足齐次或常值方差,在此情况下,对于所有的 i,有 $\sigma_i^2=\sigma^2$。于是,误差服从 iid $(0,\sigma^2)$,

称为球面误差(spherical error),因为 $\mathbf{\Omega}=\sigma^2\mathbf{I}$。

假设 5 替代了一个条件异方差回归误差(**conditional heteroskedastic regression errors**)假设,其中,异方差意味着异质性方差或不同方差。这个假设可用二阶矩 $\mathrm{E}[u^2|\mathbf{x}]=0$ 来陈述,但是由假设 4,$\mathrm{E}[u|\mathbf{x}]=0$,所以它等于方差 $\mathrm{V}[u|\mathbf{x}]$。这种更一般的异方差误差的假设,是由于它在实证上经常用于横截面回归而产生的。进一步,放松同方差性假设的代价并不高,因为即使异方差性的函数形式是未知的,获得 OLS 估计量的有效标准误差也是可能的。

使用条件异方差术语源于下述原因。即使$(y_i, \mathbf{x}_i)$是 iid 的,如同在简单随机抽样情况下,一旦我们以 $\mathbf{x}_i$ 为条件,其条件均值与条件方差都会随着 $\mathbf{x}_i$ 而变化。类似地,在简单随机抽样下,误差 $u_i=y_i-\mathbf{x}_i'\boldsymbol{\beta}$是 iid 的,因此,它们是无条件同方差的。一旦我们以 $\mathbf{x}_i$ 为条件,并且考虑以 $\mathbf{x}_i$ 为条件的 u_i 分布,就允许这个条件的方差随 $\mathbf{x}_i$ 而变化。

$N^{-1/2}\mathbf{X}'\mathbf{u}$ 的极限方差矩阵

为了获得 $N^{-1/2}\mathbf{X}'\mathbf{u}$ 的极限方差矩阵,就需要假设 6。如果回归元与误差是独立的,这就是一个比假设 4 更强的假设,那么假设 5 即 $\mathrm{E}[|u_i^2|^{1+\delta}]<\infty$和假设 3 即 $\mathrm{E}[|x_{ij}x_{ik}|^{1+\delta}]<\infty$,蕴含着假设 6 的条件即 $\mathrm{E}[|u_i^2x_{ij}x_{ik}|^{1+\delta}]<\infty$ 。

我们故意没有做出第 7 个假设:误差 $\mathbf{u}$ 是以 $\mathbf{X}$ 为条件的正态分布。为了获得 OLS 估计量的精确小样本分布,就需要譬如正态性的假设。然而,本书中自始至终关注渐近方法,因为微观经济计量学中使用的估计量极少利用精确的小样本分布结果,进而不再需要正态假设。

4.4.8 *OLS 估计量推断*

这里,我们既阐述 OLS 估计量的小样本分布以及极限分布,又在假设 1～假设 6 的条件下,验证 OLS 估计量的方差矩阵的怀特估计量。

小样本分布

在假设 1～假设 4 下,参数 $\boldsymbol{\beta}$是可识别的,从而 $\mathrm{E}[\mathbf{y}|\mathbf{X}]=\mathbf{X}\boldsymbol{\beta}$,并且 $\mathbf{X}$ 的秩为 K。

小样本中,在假设 1～假设 4 下,OLS 估计量是无偏的,其方差矩阵很容易在给定假设 5 时获得。要获得这些结果,可以利用期望迭代定理,首先对以 $\mathbf{X}$ 为条件的 $\mathbf{u}$ 取期望,然后取无条件期望。于是,由式(4.11)知:

$$\begin{aligned}\mathrm{E}[\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}]&=\boldsymbol{\beta}+\mathrm{E}_{\mathbf{X},\mathbf{u}}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}]\\&=\boldsymbol{\beta}+\mathrm{E}_{\mathbf{X}}[\mathrm{E}_{\mathbf{u}|\mathbf{X}}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|X]]\\&=\boldsymbol{\beta}+\mathrm{E}_{\mathbf{X}}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}_{\mathbf{u}|\mathbf{X}}[\mathbf{u}|\mathbf{X}]]\\&=\boldsymbol{\beta}\end{aligned}\tag{4.24}$$

利用期望迭代定理(定理 A. 23),并已知假设 1 与假设 4,这些一起推导出 $\mathrm{E}[\mathbf{u}|\mathbf{X}]=\mathbf{0}$。类似地,已知假设 5,由式(4. 11)得到:

$$\mathrm{V}[\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}]=\mathrm{E}_{\mathbf{X}}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}\,\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}]\tag{4.25}$$

其中,$\mathrm{E}[\mathbf{u}\mathbf{u}'|\mathbf{X}]=\mathbf{\Omega}$,并且我们使用定理 A. 23,可知:

$$\mathrm{V}_{\mathbf{X},\mathbf{u}}[\mathbf{g}(\mathbf{X},\ \mathbf{u})]=\mathrm{E}_{\mathbf{X}}[\mathrm{V}_{\mathbf{u}|\mathbf{X}}[\mathbf{g}(\mathbf{X},\ \mathbf{u})]]+\mathrm{V}_{\mathbf{X}}[\mathrm{E}_{\mathbf{u}|\mathbf{X}}[\mathbf{g}(\mathbf{X},\ \mathbf{u})]]$$

当 $\mathrm{E}_{\mathbf{u}|\mathbf{X}}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}]=\mathbf{0}$,即第二项为零时,得到简化式。

因此,若 $\mathrm{E}[\mathbf{u}|\mathbf{X}]=\mathbf{0}$,则 OLS 估计量是无偏的(**unbiased**)。这个有价值的性质通常不能扩展到非线性估计量上。大多数非线性估计量,比如非线性最小二乘估计量,都是有偏的,并且甚至有一些线性估计量,例如工具变量估计量,也是有偏的。OLS 估计量是非有效的(**inefficient**),因为其方差在线性无偏估计量中并不是最小的方差矩阵,除非 $\mathbf{\Omega}=\sigma^2\mathbf{I}$。尽管 OLS 的有效性损失不一定很大,但是,OLS 的非有效性却提供了进一步寻找更有效估计量譬如广义最小二乘法的动机。在以 $\mathbf{X}$ 为条件误差的其他正态性假设下,微观经济计量学应用中通常不做出该假设,OLS 估计量则是以 $\mathbf{X}$ 为条件正态分布的。

一致性

由假设 3,因为 $\mathrm{plim}\ N^{-1}\mathbf{X}'\mathbf{X}=\mathbf{M}_{\mathbf{XX}}$,所以 $\mathrm{plim}\ (N^{-1}\mathbf{X}'\mathbf{X})^{-1}=\mathbf{M}_{XX}^{-1}$。于是,一致性需要的条件(4.13)得到满足。一旦把大数定律应用于 $N^{-1}\mathbf{X}'\mathbf{u}=N^{-1}\sum_i \mathbf{x}_i u_i$,若 $\mathrm{E}[\mathbf{x}_i u_i]=\mathbf{0}$,则它依概率收敛于 0,从而这个条件被建立起来。给定假设 1 与假设 2,$\mathbf{x}_i u_i$ 是 inid 的,同时假设 1~假设 5 允许使用马尔可夫 LLN(定理 A.9)。如果假设 1 被简化成$(y_i,\ \mathbf{x}_i)$为 iid 的,那么 $\mathbf{x}_i u_i$ 是 iid 的,并且假设 1~假设 4 允许使用较简单的柯尔莫哥洛夫(Kolmogorov)LLN(定理 A.8)。

极限分布

由假设 3 知,$\mathrm{plim}\ (N^{-1}\mathbf{X}'\mathbf{X})^{-1}=\mathbf{M}_{\mathbf{XX}}^{-1}$。关键在于通过利用中心极限定理获得 $N^{-1/2}\mathbf{X}'\mathbf{u}=N^{-1/2}\sum_i \mathbf{x}_i u_i$ 的极限分布。已知假设 1 与假设 2,$\mathbf{x}_i u_i$ 是 inid 的,并且假设 1~假设 6 允许使用李雅普诺夫 CLT(定理 A.15)。如果假设 1 被加强成$(y_i,\ \mathbf{x}_i)$为 iid 的,那么 $\mathbf{x}_i u_i$ 是 iid 的,并且假设 1~假设 5 允许使用较简单的林德伯格—利维(Lindeberg-Levy) CLT(定理 A.14)。

这就得出:

$$\frac{1}{\sqrt{N}}\mathbf{X}'\mathbf{u}\xrightarrow{d}\mathcal{N}[\mathbf{0},\ \mathbf{M}_{\mathbf{X\Omega X}}] \tag{4.26}$$

其中,$\mathbf{M}_{\mathbf{X\Omega X}}=\mathrm{plim}\ N^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}=\mathrm{plim}\ N^{-}\sum_i u_i^2\ \mathbf{x}_i\mathbf{x}_i'$ 独立于 i。一旦 $\mathrm{E}_{u_i,\mathbf{x}_i}[u_i^2\mathbf{x}_i\mathbf{x}_i']=\mathrm{E}_{\mathbf{x}_i}[\mathrm{E}[u_i^2\mid\mathbf{x}_i]\mathbf{x}_i\mathbf{x}_i']$ 以及 $\sigma_i^2=\mathrm{E}[u_i^2\mid\mathbf{x}_i]$,应用大数定律得出,$\mathbf{M}_{\mathbf{X\Omega X}}=\lim\ N^{-1}\times\sum_i\mathrm{E}_{\mathbf{x}_i}[\sigma_i^2\mathbf{x}_i\mathbf{x}_i']$。由此可得,$\mathbf{M}_{\mathbf{X\Omega X}}=\mathrm{plim}\ N^{-1}\mathrm{E}[\mathbf{X}'\mathbf{\Omega}\mathbf{X}]$,其中,$\mathbf{\Omega}=\mathrm{Diag}[\sigma_i^2]$,并且期望值只与 $\mathbf{X}$ 有关,而不是与 $\mathbf{X}$ 和 $\mathbf{u}$ 都有关。

此处的阐述均假定,对于不同的 i 具有独立性。更一般地,我们允许观测值相关。于是,$\mathbf{M}_{\mathbf{X\Omega X}}=\mathrm{plim}\ N^{-1}\sum_i\sum_j u_i u_j\mathbf{x}_i\mathbf{x}_j'$,并且 $\mathbf{\Omega}$ 的第 i 行、第 j 列元素是 $\sigma_{ij}=\mathrm{Cov}[u_i,\ u_j]$。这种复杂情况将在 5.8 节的对非线性 LS 估计量研究中加以处理。

异方差性稳健的标准误差

我们考察对 $\mathbf{M}_{\mathbf{X\Omega X}}$ 进行一致估计的关键步骤。从最初的 $\mathbf{M}_{\mathbf{X\Omega X}}=\mathrm{plim}\ N^{-1}\times\sum_{i=1}^N u_i^2\mathbf{x}_i\mathbf{x}_i'$定义开始,我们用$\hat{u}_i=y_i-\mathbf{x}_i'\hat{\boldsymbol{\beta}}$ 代替 u_i,其中,由于$\hat{\boldsymbol{\beta}}\xrightarrow{p}\boldsymbol{\beta}$,所以在渐近形式上有$\hat{u}_i\xrightarrow{p}u_i$。这就得出一致估计量:

$$\hat{\mathbf{M}}_{\mathbf{x}\Omega\mathbf{x}}=\frac{1}{N}\sum_{i=1}^{N}\hat{u}_i^2\mathbf{x}_i\mathbf{x}_i'=N^{-1}\mathbf{X}'\hat{\mathbf{\Omega}}\mathbf{X} \tag{4.27}$$

其中，$\hat{\mathbf{\Omega}}=\text{Diag}[\hat{u}_i^2]$。对于正的常数 δ、Δ 以及 $j,k=1,\cdots,K$，附加假设 $\text{E}[|x_{ij}^2x_{ik}x_{il}|^{1+\delta}]<\Delta$ 是需要的，因为$\hat{u}_i^2\mathbf{x}_i\mathbf{x}_i'=(u_i-\mathbf{x}_i'(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))^2\mathbf{x}_i\mathbf{x}_i'$包括 $\mathbf{x}_i$ 的四次幂[参见怀特(White, 1980a)]。

注意到，$\hat{\mathbf{\Omega}}$ 并没有收敛到 $N\times N$ 阶矩阵 $\mathbf{\Omega}$，因为存在 N 个方差 σ_i^2 需要加以估计，从表面上看，在没有额外结构的条件下，这是不可能完成的。但是，所需要的全部内容是，$N^{-1}\mathbf{X}'\hat{\mathbf{\Omega}}\mathbf{X}$ 收敛到 $K\times K$ 阶矩阵，$\text{plim}\,N^{-1}\mathbf{X}'\mathbf{\Omega}\mathbf{X}=N^{-1}\text{plim}\sum_i\sigma_i^2\mathbf{x}_i\mathbf{x}_i'$。这比较容易得到，因为回归元 K 的个数是固定的。为了理解怀特估计量，考察只有一个截距且具有异方差误差的 $y_i=\beta+u_i$ 模型的 OLS 估计量。使用我们的记号，可以证明，$\hat{\beta}=\bar{y}$，$\mathbf{M}_{\mathbf{xx}}=\text{plim}\,N^{-1}\sum_i1=1$，并且 $\mathbf{M}_{\mathbf{x}\Omega\mathbf{x}}=\text{plim}\,N^{-1}\sum_i\text{E}[u_i^2]$，其中，$\hat{u}_i=y_i-\hat{\beta}$。$\mathbf{M}_{\mathbf{x}\Omega\mathbf{x}}$的一个明显估计量是 $\hat{\mathbf{M}}_{\mathbf{x}\Omega\mathbf{x}}=N^{-1}\sum_i\hat{u}_i^2$。为了获得这个估计量的概率极限，考察 $N^{-1}\sum_iu_i^2$ 就足够，因为给定 $\hat{\beta}\xrightarrow{p}\beta$ 时，有$\hat{u}_i-u_i\xrightarrow{p}0$。如果大数定律能用于这个平均值，收敛到它的期望值极限，那么 $\text{plim}\,N^{-1}\sum_iu_i^2=\text{plim}\,N^{-1}\sum_i\text{E}[u_i^2]=\mathbf{M}_{\mathbf{x}\Omega\mathbf{x}}$。艾克(Eicker, 1967)给出这个事例的正式条件。

4.5 加权最小二乘法

如果需要使用稳健标准误差，那么提高有效性通常是可能的。例如，如果出现异方差，那么可行广义最小二乘法(GLS)估计量就比 OLS 估计量更有效。

在本节中，我们将介绍可行 GLS 估计量，此估计量对误差项的方差做出更强的分布假设。不过，正如 OLS 情况一样，可能获得可行 GLS 估计量的标准误差对于误差方差错误设定而言，该估计量是稳定的。

在微观经济计量学中，许多研究并没有利用 GLS 的潜在有效性优势，这是由于方便性以及有效性提高相对很小。相反，普遍使用稍欠有效的加权最小二乘法，尤其是 OLS，它具有对标准误差的稳健估计。

4.5.1 *GLS 和可行 GLS*

由引论课程中阐述的高斯—马尔可夫理论可知，如果线性回归模型误差是独立且同方差的，那么 OLS 估计量是线性无偏估计量中有效的估计量。

然而，我们假定误差方差矩阵 $\mathbf{\Omega}\neq\sigma^2\mathbf{I}$。若 $\mathbf{\Omega}$ 是已知的且非奇异的，则我们用 $\mathbf{\Omega}^{-1/2}$乘以线性回归模型(4.8)，其中，$\mathbf{\Omega}^{-1/2}\mathbf{\Omega}^{1/2}=\mathbf{\Omega}$，得到：

$$\mathbf{\Omega}^{-1/2}\mathbf{y}=\mathbf{\Omega}^{-1/2}\mathbf{X}\boldsymbol{\beta}+\mathbf{\Omega}^{-1/2}\mathbf{u}$$

经过一些代数运算，得出 $\text{V}[\mathbf{\Omega}^{-1/2}\mathbf{u}]=\text{E}[(\mathbf{\Omega}^{-1/2}\mathbf{u})(\mathbf{\Omega}^{-1/2}\mathbf{u})'|\mathbf{X}]=\mathbf{I}$。因此，在这种转换模型中，误差是零均值、不相关且同方差的。因而，通过 $\mathbf{\Omega}^{-1/2}\mathbf{X}$ 对 $\mathbf{\Omega}^{-1/2}\mathbf{y}$ 的 OLS 回归，有效地估计出 $\boldsymbol{\beta}$。

这一推导得出广义最小二乘估计量(**generalized least-squares estimator**)：

$$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}=(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y} \tag{4.28}$$

GLS 估计量并不能直接实施，因为在实际应用中，$\boldsymbol{\Omega}$ 是未知的。相反，我们设定 $\boldsymbol{\Omega}=\boldsymbol{\Omega}(\boldsymbol{\gamma})$，其中，$\boldsymbol{\gamma}$ 表示有限维参数向量，获得 $\boldsymbol{\gamma}$ 的一致估计量 $\hat{\boldsymbol{\gamma}}$ 并且建立 $\hat{\boldsymbol{\Omega}}=\boldsymbol{\Omega}(\hat{\boldsymbol{\gamma}})$。例如，如果误差是异方差的，那么设定 $\mathrm{V}[u|\mathbf{x}]=\exp(\mathbf{z}'\boldsymbol{\gamma})$，其中，$\mathbf{z}$ 表示 $\mathbf{x}$ 的子集，同时使用指数函数来确保正方差。然后，$\hat{\boldsymbol{\gamma}}$ 可以通过 OLS 残差平方 $\hat{u}_i^2=(y-\mathbf{x}'\hat{\boldsymbol{\beta}}_{\mathrm{OLS}})^2$ 对 $\exp(\mathbf{z}'\boldsymbol{\gamma})$ 的非线性最小二乘法回归（参见 5.8 节）得到一致估计。此估计值可用来代替式(4.28)中的 $\boldsymbol{\Omega}$ 。注意到，我们不能用 $\hat{\boldsymbol{\Omega}}=\mathrm{Diag}[\hat{u}_i^2]$来代替式(4.28)中的 $\boldsymbol{\Omega}$，因为这会产生非一致的估计量（参见 5.8.6 节）。

可行广义最小二乘（FGLS）估计量[**feasible generalized least-squares (FGLS) estimator**]是：

$$\hat{\boldsymbol{\beta}}_{\mathrm{FGLS}}=(\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X})\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{y} \tag{4.29}$$

如果假设 1 至假设 6 满足，并且 $\boldsymbol{\Omega}(\boldsymbol{\gamma})$是正确设定的（下面将会放松这个强假设），同时 $\hat{\boldsymbol{\gamma}}$ 关于 $\boldsymbol{\gamma}$ 是一致的，那么可以证明：

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{\mathrm{FGLS}}-\boldsymbol{\beta})\xrightarrow{d}\mathcal{N}[\mathbf{0},\ (\operatorname{plim} N^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}] \tag{4.30}$$

FGLS 估计量与 GLS 估计量有相同的极限方差矩阵，从而是二阶矩有效的。实施中，用式(4.30)中的 $\hat{\boldsymbol{\Omega}}$ 代替 $\boldsymbol{\Omega}$。

可以证明，GLS 估计量最小化 $\mathbf{u}'\boldsymbol{\Omega}^{-1}\mathbf{u}$，参见习题 4.5，如果误差是异方差的且不相关的，那么 $\mathbf{u}'\boldsymbol{\Omega}^{-1}\mathbf{u}$ 可简化成$\sum_i u_i^2/\sigma_i^2$。提供 GLS 的动机是求出 $\boldsymbol{\beta}$ 的有效估计。根据 4.2 节对损失函数与最优预测的讨论，就异方差误差而言，损失函数是 $\mathrm{L}(e)=e^2/\sigma^2$。与具有 $\mathrm{L}(e)=e^2$ 的 OLS 相比，GLS 损失函数对关于具有大条件误差方差的观测值预测误差施加相对较小的惩罚。

4.5.2 *加权最小二乘法*

式(4.30)的结果假定对误差方差矩阵 $\boldsymbol{\Omega}(\boldsymbol{\gamma})$做出了正确设定。然而，若 $\boldsymbol{\Omega}(\boldsymbol{\gamma})$被错误设定，则 FGLS 估计量仍然是一致的，但是，式(4.30)给出一个错误的方差。幸运的是，即使 $\boldsymbol{\Omega}(\boldsymbol{\gamma})$被错误设定，仍然可以得到 GLS 估计量方差的一个稳健估计量。

特定地，定义 $\boldsymbol{\Sigma}=\boldsymbol{\Sigma}(\boldsymbol{\gamma})$为实用方差矩阵（**working variance matrix**），这不必等于真正的方差矩阵 $\boldsymbol{\Omega}=\mathrm{E}[\mathbf{uu}'|\mathbf{X}]$。构造一个估计量 $\hat{\boldsymbol{\Sigma}}=\boldsymbol{\Sigma}(\hat{\boldsymbol{\gamma}})$，其中，$\hat{\boldsymbol{\gamma}}$ 表示 $\boldsymbol{\gamma}$ 的估计值。然后，使用带有加权矩阵 $\hat{\boldsymbol{\Sigma}}^{-1}$的加权最小二乘法。

这就获得加权最小二乘法（WLS）估计量[**weight least-squares (WLS) estimator**]：

$$\hat{\boldsymbol{\beta}}_{\mathrm{WLS}}=(\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{y} \tag{4.31}$$

那么，统计推断在没有假设 $\boldsymbol{\Sigma}=\boldsymbol{\Omega}$ 的情况下就可进行，此处假设 $\boldsymbol{\Sigma}=\boldsymbol{\Omega}$ 是真实误差项的方差矩阵。在统计学文献中，这个方法被称为实用矩阵方法，我们称之为加权最小二乘法。但是注意到，其他一些学者却用加权最小二乘法意指 $\boldsymbol{\Omega}^{-1}$处于对角线时的 GLS 或 FGLS。这里，并没有加权矩阵 $\boldsymbol{\Sigma}^{-1}=\boldsymbol{\Omega}^{-1}$的假设。

对由 4.4.5 节给出的 OLS 进行类似的代数计算，可得到估计渐近方差矩阵：

$$\hat{\mathrm{V}}[\hat{\beta}_{\mathrm{WLS}}]=(\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Omega}}\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}(\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-1} \tag{4.32}$$

其中，$\hat{\boldsymbol{\Omega}}$ 使得：

$$\operatorname{plim} N^{-1}\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Omega}}\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}=\operatorname{plim} N^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega}\boldsymbol{\Sigma}^{-1}\mathbf{X}$$

在异方差情况下，$\hat{\boldsymbol{\Omega}}=\mathrm{Diag}[\hat{u}_i^{*2}]$，其中，$\hat{u}_i^*=y_i-\mathbf{x}_i'\hat{\beta}_{\mathrm{WLS}}$。

对于异方差误差，基本方法是选择异方差性的简单模型，例如，误差方差只依赖于一个或两个关键回归元。例如，在作为受教育与其他一些变量的函数的工资水平的线性回归模型中，异方差性可能被建模成唯一受教育的函数。假定由这个模型得出 $\hat{\boldsymbol{\Sigma}}=\mathrm{Diag}[\hat{\sigma}_i]$。然后，由 $y_i/\hat{\sigma}_i$ 对 $x_i/\hat{\sigma}_i$ 的 OLS 回归（没有常值选项）得出 $\hat{\beta}_{\mathrm{WLS}}$，同时可以证明，来自此回归的怀特稳健标准误差等于建立在式(4.32)基础上的怀特稳健标准误差。

当存在不止一个复杂因素时，加权最小二乘法或者实用矩阵方法是特别方便的。例如，第 21 章的随机效应面板数据模型中，误差可被处理成针对给定个体是、与时间相关的，并且是异方差的。人们使用随机效应估计量，此估计量只能控制第一个复杂因素，另一方面要计算此估计量的异方差一致标准误差。

各种最小二乘估计量已总结在表 4.2 中。

表 4.2　最小二乘估计量和它们的渐近方差

估计量[a]	定　义	估计量渐近方差
OLS	$\hat{\beta}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$	$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\boldsymbol{\Omega}}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$
FGLS	$\hat{\beta}=(\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{y}$	$(\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X})^{-1}$
WLS	$\hat{\beta}=(\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{y}$	$(\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Omega}}\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}(\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-1}$

[a] 估计量是具有误差条件方差矩阵 Ω 的线性回归模型估计量。对 FGLS 来说，假设 $\hat{\Omega}$ 关于 Ω 是一致的。对 OLS 与 WLS 来说，$\hat{\beta}$ 的异方差稳健方差使用 $\hat{\Omega}$，$\hat{\Omega}$ 等于一个对角具有平方残差的对角矩阵。

4.5.3　*LS 稳健标准误差的事例*

举一个稳健标准误差估计量的事例，考虑下面关于具有乘法异质性 dgp 的斜率系数最小二乘估计的标准误差的估计：

$$y=1+1\times x+u$$
$$u=x\varepsilon$$

其中，纯量回归元 $x\sim\mathcal{N}[0,\ 25]$，而 $\varepsilon\sim\mathcal{N}[0,\ 4]$。

误差是条件异方差的，因为 $\mathrm{V}[u\mid x]=\mathrm{V}[x\varepsilon\mid x]=x^2\mathrm{V}[\varepsilon\mid x]=4x^2$，它依赖于回归元 x。给定 x，ε 是独立的，此处的 dgp 是特定的，这不同于无条件方差，其中，$\mathrm{V}[u]=\mathrm{V}[x\varepsilon]=\mathrm{E}[(x\varepsilon)^2]-(\mathrm{E}[x\varepsilon])^2=\mathrm{E}[x^2]\mathrm{E}[\varepsilon^2]=\mathrm{V}[x]\mathrm{V}[\varepsilon]=100$。

OLS 估计量的标准误差，应该利用异方差一致的或稳健方差估计(4.21)来进行计算。由于 OLS 不是完全有效的，所以 WLS 可能促使有效性提高。GLS 将肯

定促使有效性提高。并且，在这个模拟数据的例子中，我们知道，$V[u|x]=4x^2$。所有估计方法均得出截距与斜率参数的一致估计。

来自容量为 100 的样本生成数据，各种不同最小二乘法估计和与之有关的标准误差都已由表 4.3 给出。我们考虑斜率系数。

表 4.3　最小二乘法带有条件异方差误差的例子[a]

	OLS	WLS	GLS
常数	2.213	1.060	0.996
	(0.823)	(0.150)	(0.007)
	[0.820]	[0.051]	[0.006]
x	0.979	0.957	0.952
	(0.178)	(0.190)	(0.209)
	[0.275]	[0.232]	[0.208]
R^2	0.236	0.205	0.174

[a] 对容量为 100 的样本生成的数据。OLS、WLS 以及 GLS 全部是一致的，但 OLS 与 WLS 却是非常有效的。给出两种不同的标准误差：圆括号中的是假设同方差误差的默认标准误差；方括号中的是异方差稳健标准误差。数据生成过程在下一节给出。

OLS 斜率系数估计值是 0.979。两个标准误差估计值均已报告出来，利用式(4.21)获得的正确异方差性稳健标准误差为 0.275，它比利用 $s^2(\mathbf{X}'\mathbf{X})^{-1}$ 计算出的不正确估计值 0.177 要大很多。标准误差估计中如此大的差异，可以导致截然不同的统计推断结论。一般而言，标准误差的偏倚可以朝向任何方向。例如，在理论上我们可以证明，稳健标准误差的极限比不正确的极限大$\sqrt{3}$倍。特别地，对于这个 dgp 以及在样本量为 N 的情况下，斜率系数的 OLS 估计量的正确与不正确的标准误差分别收敛到$\sqrt{12/N}$与$\sqrt{4/N}$。

举一个 WLS 估计量的事例，假定 $u=\sqrt{|x|}\varepsilon$ 而不是 $u=x\varepsilon$，因此，$V[u]=\sigma^2|x|$。一旦用 y 除以截距且用 x 除以$\sqrt{x}$，WLS 估计量能够利用 OLS 回归计算出。由于这是关于异方差误差的错误模型，所以斜率系数的正确标准误差是利用式(4.32)计算出的稳健估计值 0.232。

一旦利用 y 除以截距，且 x 除以 $|x|$，可通过 OLS 回归计算出这个 dgp 的 GLS 估计量，因为变换误差是同方差的。斜率系数的通常标准误差与稳健标准误差大小差不多(0.209 与 0.208)。这是我们所希望的结果，两个值在渐近形式上都是正确的，因为此处 GLS 估计量使用了异方差性正确模型。理论上可以证明，对于这个 dgp 来说，斜率系数的 GLS 估计值的标准误差收敛到$\sqrt{4/N}$。

正如人们所预料的，OLS 与 WLS 都不如 GLS 有效，它们的斜率系数标准误差关系为 0.275>0.232>0.208。

这个事例的设置是在横截面数据估计理论中经常使用的标准设置。y 是随机的随机变量，$\mathbf{x}$ 也是随机的随机变量。$(y_i,\ \mathbf{x}_i)$对于不同 i 来说是独立的，且为同分布，如同随机抽样时的情况。然而，$y_i|\mathbf{x}_i$ 的条件分布对于不同 i 是不一样的，因为 y_i的条件均值与方差都依赖于 $\mathbf{x}_i$。

4.6 中位数与分位数回归

在只有一个截距的模型中，关于样本分布的概述统计量，除了样本均值之外，还包括分位数，譬如中位数、上四分位数、下四分位数和百分位数。

在回归背景下，类似地，我们对条件分位数感兴趣。例如，关注内容在于，与那些受教育程度高的工人相比，受教育程度低的工人薪水分布百分位数如何置于更小的空间中。在这个简单事例中，人们可以分别计算受教育程度低的工人与受教育程度高的工人的情况。然而，如果存在几个回归元取几个值的情况，这种方法就行不通。相反，为了估计给定 **x** 时 y 条件分布的分位数，就需要分位数回归。

由表 4.1 知，分位数回归对应于使用非对称的绝对损失，而中位数回归作为特殊情况，则使用绝对误差损失。这些方法提供了对 OLS 而言可供选择的方法，它们都使用误差平方的损失。

分位数回归方法除了提供数据的更丰富特性，还具有其他优点。与最小二乘法回归相比，中位数回归对离群值而言更加稳健。此外，与最小二乘法估计所需要的假设相比，分位数回归估计量在较弱的随机假设下可以是一致的。重要的事例是曼斯基(Manski，1975)关于二值结果模型的最大得分估计量(参见 14.6 节)，以及鲍威尔(Powell，1984)关于删失模型的删失最小绝对偏差估计量(参见 16.9.2 节)。

在转向样本分位数估计之前，我们以对总体分位数给出简略解释开始。

4.6.1 *总体分位数*

对于连续随机变量 y 来说，其总体第 q 分位数是 μ_q，使得 y 以概率 q 小于或等于 μ_q。因而有：

$$q=\Pr[y\leqslant\mu_q]=F_y(\mu_q)$$

其中，F_y 表示 y 的累积分布函数(cdf)。例如，如果 $\mu_{0.75}=3$，那么 $y<3$ 的概率等于 0.75。由此可得：

$$\mu_q=F_y^{-1}(q)$$

重要事例是中位数 $q=0.5$、上四分位数 $q=0.75$、下四分位数 $q=0.25$。对于标准正态分布，有 $\mu_{0.5}=0.0$、$\mu_{0.95}=1.645$ 以及 $\mu_{0.975}=1.960$。第 $100q$ 个百分位数(**percentile**)是 q 分位数。

对于回归模型，以 **x** 为条件的 y 的总体第 q 个分位数是函数，使得以 **x** 为条件的 y 以概率 q 小于或等于 $\mu_q(x)$，其概率是利用给定 **x** 时 y 的条件分布计算得到的。由此可得：

$$\mu_q(x)=F_{y|\mathbf{x}}^{-1}(q) \tag{4.33}$$

其中，$F_{y|\mathbf{x}}$ 表示给定 **x** 时 y 的条件 cdf，并且我们没有表示此分布的参数作用。

一种深刻认识是，在下述假设下去推导分位数函数 $\mu_q(\mathbf{x})$，即假定 dgp 是含有

乘法异方差性的线性模型：

$$y=\mathbf{x}'\boldsymbol{\beta}+u$$
$$u=\mathbf{x}'\boldsymbol{\alpha}\times\varepsilon$$
$$\varepsilon\sim\text{iid}\,[0,\sigma^2]$$

其中，假定 $\mathbf{x}'\boldsymbol{\alpha}>0$。于是，以 $\mathbf{x}$ 为条件的 y 的总体 q 分位数，就是使得

$$\begin{aligned}q&=\Pr[y\leqslant\mu_q(\mathbf{x},\boldsymbol{\beta},\boldsymbol{\alpha})]\\&=\Pr[u\leqslant\mu_q(\mathbf{x},\boldsymbol{\beta},\boldsymbol{\alpha})-\mathbf{x}'\boldsymbol{\beta}]\\&=\Pr[\varepsilon\leqslant[\mu_q(\mathbf{x},\boldsymbol{\beta},\boldsymbol{\alpha})-\mathbf{x}'\boldsymbol{\beta}]/\mathbf{x}'\boldsymbol{\alpha}]\\&=F_\varepsilon([\mu_q(\mathbf{x},\boldsymbol{\beta},\boldsymbol{\alpha})-\mathbf{x}'\boldsymbol{\beta}]/\mathbf{x}'\boldsymbol{\alpha})\end{aligned}$$

的那个函数 $\mu_q(\mathbf{x},\boldsymbol{\beta},\boldsymbol{\alpha})$，其中，我们使用 $u=y-\mathbf{x}'\boldsymbol{\beta}$、$\varepsilon=u/\mathbf{x}'\boldsymbol{\alpha}$ 以及 F_ε 表示 ε 的 cdf。由此可得，$[\mu_q(\mathbf{x},\boldsymbol{\beta},\boldsymbol{\alpha})-\mathbf{x}'\boldsymbol{\beta}]/\mathbf{x}'\boldsymbol{\alpha}=F_\varepsilon^{-1}(q)$，所以有：

$$\begin{aligned}\mu_q(\mathbf{x},\boldsymbol{\beta},\boldsymbol{\alpha})&=\mathbf{x}'\boldsymbol{\beta}+\mathbf{x}'\boldsymbol{\alpha}\times F_\varepsilon^{-1}(q)\\&=\mathbf{x}'(\boldsymbol{\beta}+\boldsymbol{\alpha}\times F_\varepsilon^{-1}(q))\end{aligned}$$

因而，对于含有乘法异方差性形式 $u=\mathbf{x}'\boldsymbol{\alpha}\times\varepsilon$ 的线性模型来说，条件分位数关于 $\mathbf{x}$ 是线性的。在同方差性的特殊情况下，$\mathbf{x}'\boldsymbol{\alpha}$ 等于常值，并且所有条件分位数具有相同斜率，只是它们的截距不同，截距会随着 q 增大而变大。

在更一般的事例中，分位数函数关于 $\mathbf{x}$ 可能是非线性的，其原因在于异方差性具有其他形式，比如 $u=h(\mathbf{x},\boldsymbol{\alpha})$，其中，$h(\cdot)$关于 $\mathbf{x}$ 是非线性的，或者因为回归函数本身就具有非线性形式 $g(\mathbf{x},\boldsymbol{\beta})$。在下一节，对于由式(4.34)给出的分位数回归损失函数，标准方法仍是估计线性的分位数函数，然后把它们解释成最佳线性预测量。

4.6.2 *样本分位数*

对于单变量随机变量 y，获得样本分位数估计值的通常方法是首先对样本加以排序。然后，$\hat{\mu}_q$ 等于第 $[Nq]$ 个最小值，其中，N 表示样本量，而$[Nq]$表示最接近 Nq 的最大整数。例如，若 $N=97$，则下四分位数是第 25 个观测值，因为$[97\times 0.25]=[24.25]=25$。

凯恩克和巴西特(Koenker and Bassett，1978)发现，样本的第 q 个分位数(**sample *q*th quantile**)$\hat{\mu}_q$ 能等价地表述成

$$\sum_{i:y_i\geqslant\beta}^{N}q\,|\,y_i-\beta|+\sum_{i:y_i<\beta}^{N}(1-q)\,|\,y_i-\beta|$$

求关于 β 的最小值最优化问题的解。这一结论并不明显。为了获得某种认识，考察中位数，其中，$q=0.5$。于是，中位数是 $\sum_i|\,y_i-\beta|$ 的最小值。假定在 99 个观测值的样本中，第 50 个最小观测值等于 10，即中位数，同时第 51 个最小观测值等于 12。如果我们令 $\beta=12$，而不是 10，那么对前面 50 个有序观测值来说，将增加 2；而对剩余 49 个观测值而言，将减少 2。因此，与第 50 个观测值相比，第 51 个最小观测值是一个较差的选择。类似地，可以证明，与第 50 个观测值相比，第 49 个最小

观测值是一个较差的选择。

然后，将目标函数推广到线性回归情况，因此，第 q 个分位数回归估计量（**quantile regression estimator**）$\hat{\beta}_q$ 为最小化

$$Q_N(\beta_q) = \sum_{i:y_i \geq \mathbf{x}_i'\beta}^{N} q\,|y_i - \mathbf{x}_i'\beta_q| + \sum_{i:y_i < \mathbf{x}_i'\beta}^{N} (1-q)\,|y_i - \mathbf{x}_i'\beta_q| \qquad (4.34)$$

的 β_q 值，其中，我们使用 β_q 而不是 β，以便于用 q 的各种不同选取值来估计 β 的不同值。注意，这是由表 4.1 给出的非对称绝对损失函数，其中，$\hat{y}$ 被限制成关于 $\mathbf{x}$ 是线性的，所以 $e=y-\mathbf{x}'\beta_q$。特殊情况下，$q=0.5$ 称为中位数回归估计量（**median regression estimator**）或者最小绝对偏差估计量（**least absolute deviations estimator**）。

4.6.3 *分位数回归估计量的性质*

目标函数(4.34)是不可微的，因而不能利用第 10 章阐述的梯度最优化方法。幸运的是，能使用线性规划分法，而且这些方法用于相对快速计算 $\hat{\beta}_q$。

由于 $\hat{\beta}_q$ 不存在显式解，所以不能利用 4.4 节中的 OLS 方法来获得 $\hat{\beta}_q$ 的渐近分布。因为目标函数是不可微的，同样需要对第 5 章的方法加以改进。可以证明：

$$\sqrt{N}(\hat{\beta}_q - \beta_q) \xrightarrow{d} \mathcal{N}[\mathbf{0},\ \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}] \qquad (4.35)$$

[例如，参见布钦斯基(Buchinsky, 1998, 第 85 页)]，其中：

$$\mathbf{A} = \text{plim}\,\frac{1}{N}\sum_{i=1}^{N} f_{u_q}(0\,|\,\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i' \qquad (4.36)$$

$$\mathbf{B} = \text{plim}\,\frac{1}{N}\sum_{i=1}^{N} q(1-q)\mathbf{x}_i\mathbf{x}_i'$$

而 $f_{u_q}(0|\mathbf{x})$ 表示误差项 $u_q=y-\mathbf{x}'\beta_q$ 的条件密度在 $u_q=0$ 处的计算值。$\hat{\beta}_q$ 的方差估计显得很复杂，因为需要估计 $f_{u_q}(0|\mathbf{x})$。相反，利用第 11 章的成对自助法，比较容易获得 $\hat{\beta}_q$ 的标准误差。

4.6.4 *分位数例子*

在这一节，我们实施条件分位数估计，同时将它与利用 OLS 回归的通常的条件均值估计进行比较。应用事例涉及对家庭每年医疗支出的恩格尔曲线加以估计。更具体地，我们考察医疗支出的对数与家庭总收入对数之间的回归关系。该回归产生医疗支出关于总支出常数(弹性)的估计。

数据取自世界银行的“1997 年越南生活标准调查”。样本由 5 006 个家庭组成，为了允许采取自然对数形式，在省略了 16.6%的零支出样本后，其余的家庭具有正的医疗支出水平。零值能利用 16.9.2 节中阐述的鲍威尔关于删失分位数回归方法来处理。为了简单起见，我们直接省略零支出的观测值。尤其是在低收入水平上，医疗支出的最大分量是由药店购买医疗器械构成的。尽管家庭的几个特征变量是可以利用的，但是，为了简单起见，我们只考察单一回归元，即以家庭总支出的对数作为家庭收入的代表。

线性最小二乘法回归得出 0.57 这一弹性估计值。通常，这个估计值意味着药品是“必需的”，因而对药品的需求是收入无弹性的。这种估计值并不令人感到非常惊讶，只是我们应该承认，在各种不同收入层面上，弹性存在着相当大的异质性。

正如凯恩克和哈洛克(Koenker and Hallock，2001)所强调的，分位数回归是研究这类异质性的有力工具。我们对式(4.34)求最小值，其中，y 表示医疗支出的对数，而 $\mathbf{x}'\boldsymbol{\beta}=\beta_1+\beta_2 x$，此处，$x$ 表示家庭总支出的对数。对于 19 个分位数 $q=\{0.05, 0.10, \cdots, 0.95\}$ 值都进行这样的计算，其中，$q=0.5$ 为中位数。在每种情况下，标准误差可利用含有 50 次重复抽样的自助法加以估计。该方法的结果已被归纳为图 4.1 与图 4.2。

图 4.1 中画出了取各种不同 q 值时 $\hat{\beta}_{2,q}$ 的斜率系数，以及相关的 95%置信区间。图中显示，弹性的分位数估计值是如何随分位数值而变化的。弹性估计值会系统地随家庭收入水平而增大：从 $q=0.05$ 时的 0.15 增长到 $q=0.85$ 时的 0.80 这一最大值。同样地，对最小二乘法斜率估计值 0.57 加以阐述，它作为水平线并没有随分位数而变化。显然，在较小分位数与较大分位数上的弹性估计值都是在统计上显著不同的，同时与 OLS 估计值相比也在统计上显著不同。而 OLS 具有标准误差 0.032。看起来，总弹性估计值将会依照基本收入分布的变化而变化。此图支持了由凯恩克和哈洛克引述的莫斯特勒和图基(Mosteller and Tukey，1977，第 236 页)的发现，即一旦仅仅关注于条件均值函数，最小二乘法回归将给出因变量与解释变量联合分布的不完全概括。

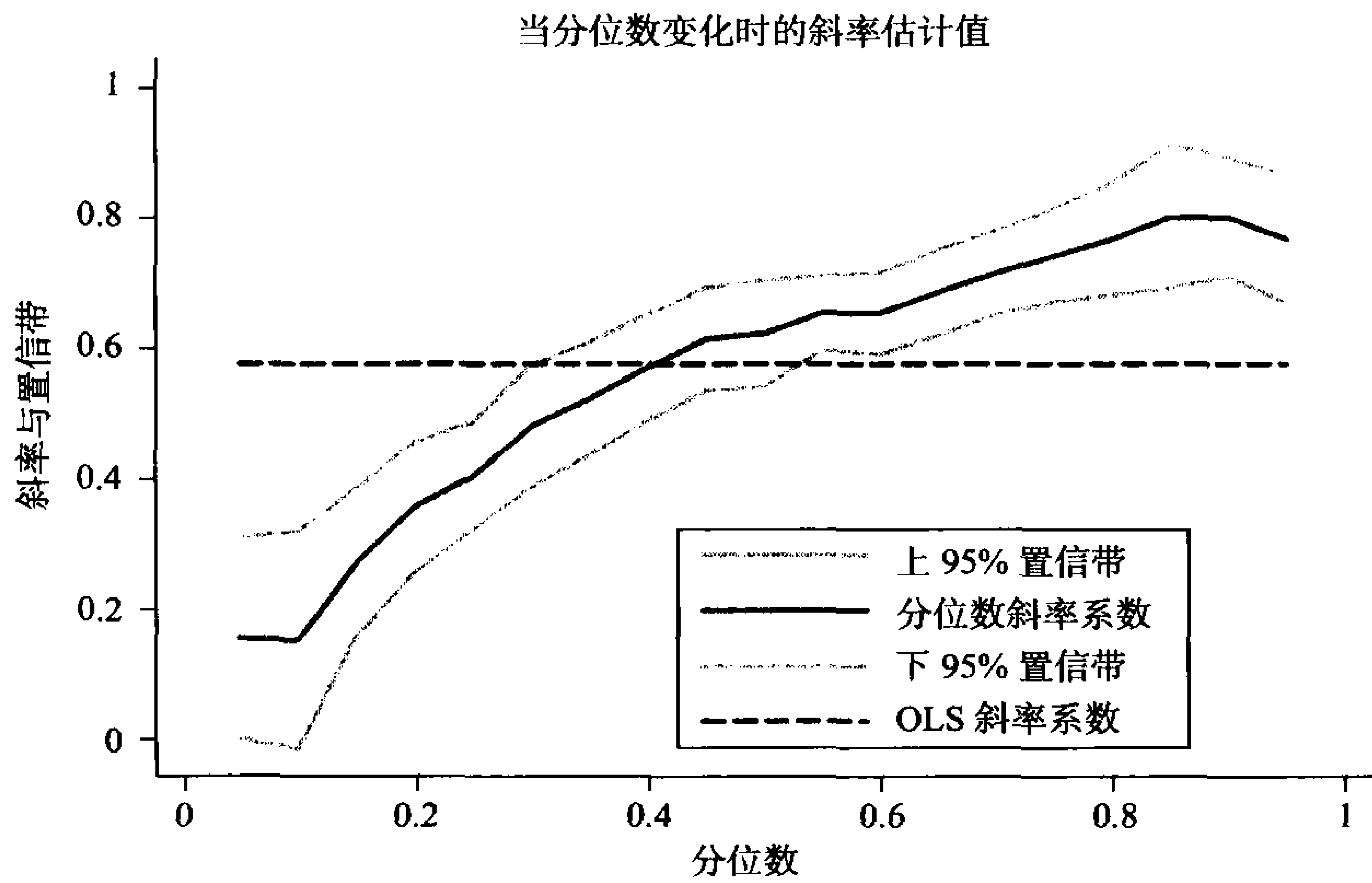


图 4.1 由医疗支出的自然对数对总支出自然对数进行回归，画出对应于 $q=0.05, 0.10, \cdots, 0.90, 0.95$ 的斜率系数的分位数回归估计值，以及相关的 95%置信带。

图 4.2 将 $q=0.1$、$q=0.5$ 和 $q=0.9$ 的三个估计分位数回归线 $\hat{y}_q=\hat{\beta}_{1,q}+\hat{\beta}_{2,q}x$ 用 OLS 回归线画在一起。没有画出 OLS 回归，它类似于中位数($q=0.5$)回归线。如图 4.2 所示，中位数回归展成扇形。并不令人感到惊讶的是，给定估计斜率随 q 而增大，正如图 4.1 所证实的。凯恩克和巴西特(Koenker and Bassett，1982)曾提

出，当数据生成过程(dgp)为线性模型时，将分位数回归作为检验异方差误差的工具。就这类情况而言，分位数回归线展成扇形可被解释成异方差性存在的证据。另一种解释是，条件均值关于 x 是线性的，并且具有递增的斜率，从而导致分位数斜率系数随分位数而增大。

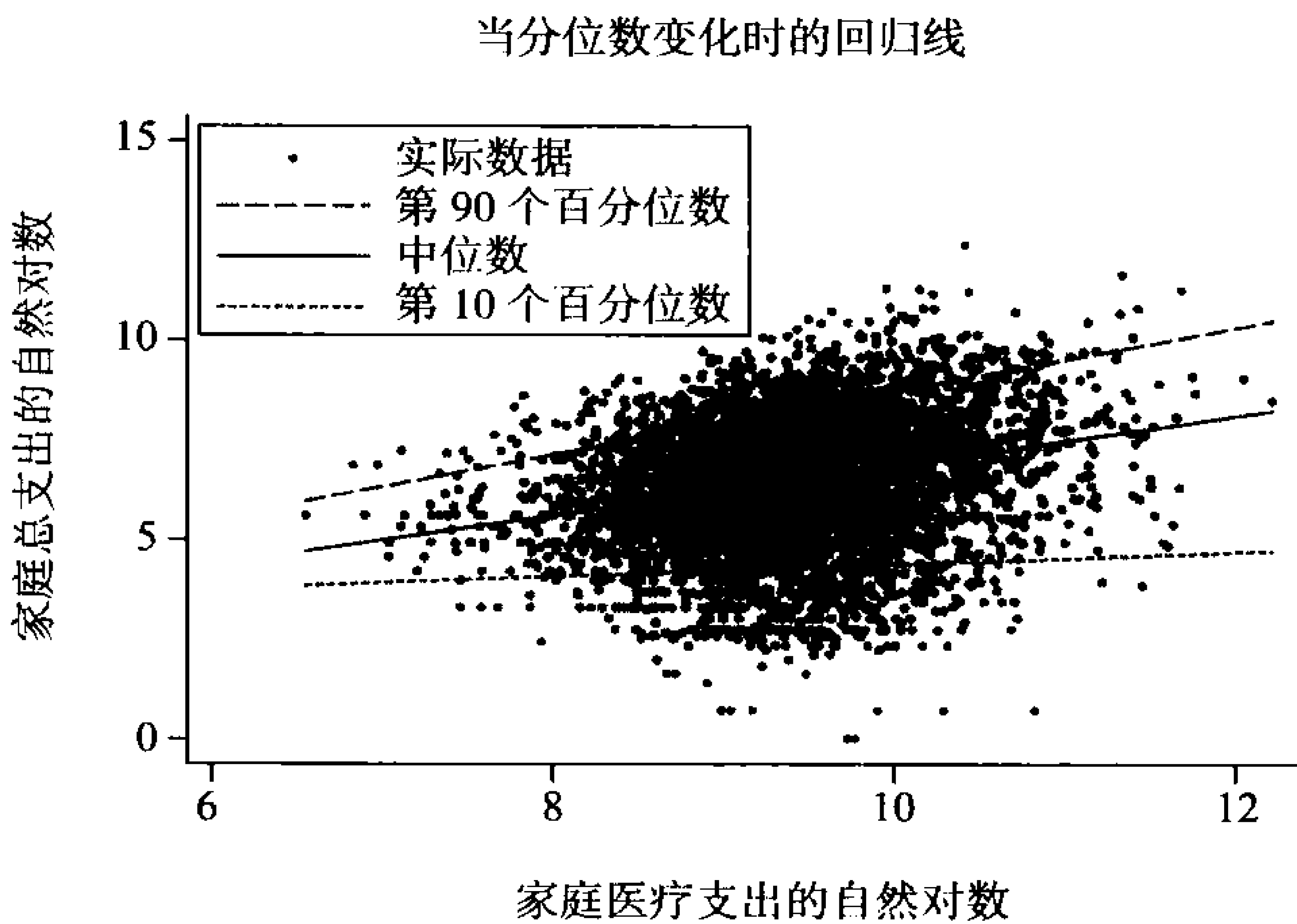


图 4.2 源自医疗支出的自然对数对总支出自然对数的回归，关于 $q=01$、$q=0.5$ 以及 $q=0.9$ 的分位数回归估计线。数据源于 1997 年具有正医疗支出的越南 5 006 个家庭数据。

有关分位数回归的更详细解释，已由布钦斯基(Buchinsky, 1994)以及凯恩克和哈洛克(Koenker and Hallock, 2001)给出。

4.7 模型错误设定

“模型错误设定”术语在最宽泛的意义上是指，对数据生成过程所做出的一个或多个不正确假设。错误设定可能单独发生，也可能联合发生，但是，如果只考察单个错误设定的结果，分析起来就比较简单。

在下面的讨论中，我们强调错误设定可能导致最小二乘法的非一致性以及所关注系数识别性的损失。然而，最小二乘法估计量可能继续拥有解释意义，与正确模型设定假设下所预期的有所不同。具体地讲，估计量可能收敛到不同于真实总体的参数上，譬如 4.7.5 节定义的伪真实值(pseudo-true value)。

这里的 OLS 一致性所引发的问题与其他模型的估计量是相关的。于是，与 OLS 一致性所需要的那些假设相比，此处的一致性需要更强一些的假设条件，所以在模型错误设定下得到的非一致性更为常见。

4.7.1 *OLS 的不一致性*

模型错误设定的最严重后果，是回归元系数 $\boldsymbol{\beta}$ 的非一致估计。由 4.4 节知，为了证明 OLS 估计量的一致性，需要两个关键性条件：(1) 数据生成过程是 $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{u}$；(2) 数据生成过程满足 $\text{plim}\, N^{-1}\mathbf{X}'\mathbf{u}=\mathbf{0}$。于是有：

$$\hat{\boldsymbol{\beta}}_{\text{OLS}}=\boldsymbol{\beta}+(N^{-1}\mathbf{X}'\mathbf{X})^{-1}N^{-1}\mathbf{X}'\mathbf{u} \\ \xrightarrow{p}\boldsymbol{\beta} \tag{4.37}$$

其中,如果 $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{u}$,那么第一个等式成立[参见式(4.12)],而第二个等式通过使用 $\operatorname{plim} N^{-1}\mathbf{X}'\mathbf{u}=\mathbf{0}$ 而得到。

若模型错误设定,从而导致关于 **y** 的错误模型(这会违背第 1 个条件),或者导致回归元与误差项相关(这会违背第 2 个条件),则 OLS 估计可能是非一致的(**inconsistent**)。

4.7.2 *函数形式错误设定*

在不确定的维数参数空间中,条件均值函数的线性设定只能以 R^K 近似真实未知条件均值函数。即使所选择的回归元正确,也可能出现条件均值被错误设定。

假定数据生成过程具有非线性回归函数的形式:

$$y=g(\mathbf{x})+v$$

其中,没有使用 $g(\mathbf{x})$ 对未知参数的相依性,同时假定 $\mathrm{E}[v|\mathbf{x}]=0$。线性回归模型:

$$y=\mathbf{x}'\boldsymbol{\beta}+u$$

是错误设定的。问题是,即使数据生成过程实际上是非线性的,OLS 估计量能否给出任何有意义的解释?

通常对回归系数解释的方法是通过真实的微观关系(**micro relationship**)来进行的,这里的微观关系为:

$$\mathrm{E}[y_i|\mathbf{x}_i]=g(\mathbf{x}_i)$$

在这种情况下,$\hat{\boldsymbol{\beta}}_{\text{OLS}}$无法测算出 $\mathrm{E}[y_i|\mathbf{x}_i]$对于 $\mathbf{x}_i$ 变化的微小响应,因为它没有收敛到$\partial g(\mathbf{x}_i)/\partial\mathbf{x}_i$。因此,不可能拥有对 $\hat{\boldsymbol{\beta}}_{\text{OLS}}$的通常解释。

怀特(White, 1980b)已经证明,OLS 估计量收敛到 $\boldsymbol{\beta}$ 值,该值使得均方预测误差

$$\mathrm{E}_{\mathbf{x}}[(g(\mathbf{x})-\mathbf{x}'\boldsymbol{\beta})^2]$$

最小化。因此,若均方预测误差用于损失函数,则由 OLS 得到的预测是非线性回归函数的最佳线性预测量。这种有用的性质已在 4.2.3 节陈述过,但是,那里没有对 $\hat{\boldsymbol{\beta}}_{\text{OLS}}$给出过多的解释。

概括地说,如果真实回归函数是非线性的,那么对于个体预测来说,OLS 作用就不大了。就预测总变化而言,OLS 仍然是有用的,它给出归因于 x 变动而引起的样本均值 $\mathrm{E}[y|\mathbf{x}]$的变化[参见斯托克(Stocker,1982)]。然而,微观经济计量分析通常探寻个体层面上有意义的模型。

本书大部分阐述关于很可能正确设定的线性模型的一些其他可供选择的方法。例如,第 14 章对二值结果的阐述,能够确保预测概率在 0 与 1 之间的模型设定。此外,人们偏爱依赖于最小分布假设的模型与方法,因为它们被错误设定的可能性很小。

4.7.3 内生性

内生性已经在 2.3 节正式定义过。内生性的一种宽泛定义是指，当回归元与误差项相关时，则该回归元是内生的。如果任何一个回归元都是内生的，那么通常所有系数的 OLS 都是非一致的(除非外生回归元与内生回归元是不相关的)。

内生性的一些重要事例包括联立方程偏倚(2.4 节)、省略变量偏倚(4.7.4 节)、样本选择偏倚(16.5 节)以及测量误差偏倚(第 26 章)，本书在线性模型和非线性模型的背景下对这些内容有着广泛研究和应用。当使用横截面观测数据时，极有可能发生内生性，而且经济学家也非常关注此类复杂情况。

控制内生性的相当一般的方法是工具变量法，这将在 4.8 节、4.9 节、6.4 节以及 6.5 节加以阐述。然而，当没有必需的工具可以利用时，这种方法就不能得到应用。

控制内生性的其他方法已经在 2.8 节阐述过，包括控制混合变量法；若有重复横截面数据或面板数据可以利用，则运用差异中差分(参见第 21 章)；若有面板数据可以利用并且内生性产生于时常值的省略变量，则运用固定效应(参见 21.6 节)，以及回归不连续(regression-discontinuity)设计(参见 25.6 节)。

4.7.4 省略变量

在引论课程中，线性回归方程的省略变量(omission of a variable)是被经常阐述的 OLS 非一致性的第一个事例。这种省略可能来自错误排除那些可以利用数据的变量，也可能来自排除那些不能被直观观测的变量。例如，在工资(或者更经常的是工资对数)对受教育的回归中忽略能力，这经常归因于能力综合测量的不可利用性。

设真实 dgp 是：

$$y=\mathbf{x}'\boldsymbol{\beta}+z\alpha+v \tag{4.38}$$

其中，$\mathbf{x}$ 与 z 均表示回归元，为了简单起见，这里，z 表示纯量回归元，而 v 表示误差项，且假定 v 与回归元 $\mathbf{x}$ 及 z 是不相关的。y 对 $\mathbf{x}$ 及 z 的 OLS 回归，将产生 $\boldsymbol{\beta}$ 与 α 的一致参数估计值。

相反，假定 y 只单独地对 $\mathbf{x}$ 进行回归，归因于不可利用性而省略了 z。于是，$z\alpha$ 项被并入误差项。估计模型是：

$$y=\mathbf{x}'\boldsymbol{\beta}+(z\alpha+v) \tag{4.39}$$

其中，误差项是 $(z\alpha+v)$。像以往一样，v 与 $\mathbf{x}$ 是不相关的，但是，如果 z 与 $\mathbf{x}$ 是相关的，那么误差项 $(z\alpha+v)$ 将与回归元 $\mathbf{x}$ 相关。OLS 估计量关于 $\boldsymbol{\beta}$ 将是非一致的。

这个模型有足够的结构来决定非一致性的方向。一旦以一种明显方式对所有观测值进行叠放，则得到 $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{z}\alpha+\mathbf{v}$。将其代入 $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$，得出：

$$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}=\boldsymbol{\beta}+(N^{-1}\mathbf{X}'\mathbf{X})^{-1}(N^{-1}\mathbf{X}'\mathbf{z})\alpha+(N^{-1}\mathbf{X}'\mathbf{X})^{-1}(N^{-1}\mathbf{X}'\mathbf{v})$$

在 $\mathbf{X}$ 与 $\mathbf{v}$ 不相关的通常假设下，最后一项概率极限为 0；然而，$\mathbf{X}$ 与 $\mathbf{z}$ 是相关的，并且有：

$$\text{plim}\ \hat{\boldsymbol{\beta}}_{\text{OLS}}=\boldsymbol{\beta}+\delta\alpha \tag{4.40}$$

其中：

$$\delta=\text{plim}[(N^{-1}\mathbf{X}'\mathbf{X})^{-1}(N^{-1}\mathbf{X}'\mathbf{z})]$$

表示省略回归元($\mathbf{z}$)对所包含回归元($\mathbf{X}$)进行回归的 OLS 估计量的概率极限。

这种非一致性称为省略变量偏倚(**omitted variables bias**)，虽然各种错误设定在形式上都会导致非一致性，但是这一普遍术语可以表述成：各种错误设定导致偏倚。只要 $\delta\neq 0$，也就是说，只要省略变量与所包含的回归元是相关的，就存在非一致性。通常，非一致性可能是正的，也可能是负的，并且其符号甚至可能与 OLS 系数的符号相反。

对于受教育事例来说，可以认为，受教育与能力之间的相关系数为正，所以 $\delta>0$，进而认为能力回报也为正，因而 $\alpha>0$。由此可得，$\delta\alpha>0$，因此，在这个事例中，省略变量偏倚是正的。工资仅对受教育的 OLS 将会高估教育在工资的影响。

错误设定的有关形式包含不相干回归元(**inclusion of irrelevant regressors**)。例如，虽然数据生成过程是更为简单的 $y=\mathbf{x}'\boldsymbol{\beta}+v$，然而，回归可能是 y 对 $\mathbf{x}$ 与 z 的回归。在这种情况下，可直接证明，该 OLS 是一致的，却损失了有效性。

如果参数估计是要给出因果解释，就必须控制省略变量偏倚。因为太多的回归元不会有很大的影响，而太少的回归元可能导致非一致性，所以由大数据集估计的微观经济计量学模型倾向于包括众多回归元。如果仍要阐述省略变量，就需要用到在 4.7.3 节末尾处给出的方法。

4.7.5 伪真实值

在省略变量事例中，最小二乘估计量受限于在混杂(confounding)意义下不能估计 $\boldsymbol{\beta}$，却可以估计 $\boldsymbol{\beta}$、δ 和 α 的函数。

OLS 估计量不能用于估计 $\boldsymbol{\beta}$。例如，它可以测算回归元 $\mathbf{x}$ 外生变化的影响，例如，在保持包括能力常值的所有其他回归元不变时。

然而，由式(4.40)知，$\hat{\boldsymbol{\beta}}_{\text{OLS}}$是函数$(\boldsymbol{\beta}+\delta\alpha)$的一致估计量，而且具有有意义的解释。$\boldsymbol{\beta}^*=\boldsymbol{\beta}+\delta\alpha$ 的 OLS 估计量 $\hat{\boldsymbol{\beta}}_{\text{OLS}}$的概率极限，称为对应于 $\hat{\boldsymbol{\beta}}_{\text{OLS}}$的伪真实值(**pseudo-true value**)，参见 5.7.1 节的正式定义。

进一步地，人们能得到$\hat{\boldsymbol{\beta}}_{\text{OLS}}$的分布，即使它关于 $\boldsymbol{\beta}$是非一致的。$\hat{\boldsymbol{\beta}}_{\text{OLS}}$的估计渐近方差测算了围绕 $(\boldsymbol{\beta}+\boldsymbol{\delta}\alpha)$的离差，并且如果式(4.38)中的误差是同方差的，那么它可由普通估计量譬如 $s^2(\mathbf{X}'\mathbf{X})^{-1}$来加以估计。

4.7.6 参数多样性

到目前为止，我们允许回归元与误差项随不同个体而变化，却把回归参数 $\boldsymbol{\beta}$限定为对不同个体而言是相同的。

相反，可以假定数据生成过程是：

$$y_i = \mathbf{x}_i'\boldsymbol{\beta}_i + u_i \tag{4.41}$$

其中，参数 $\boldsymbol{\beta}_i$ 具有下标 i。这是一个参数异质性的事例，其边际效用 $\mathrm{E}[y_i|\mathbf{x}_i]=\boldsymbol{\beta}_i$ 现在允许随不同个体而变化。

随机系数模型（**random coefficients model**）或者随机参数模型（**random parameters model**）均把 $\boldsymbol{\beta}_i$ 设定为独立同分布的，该分布不依赖于观测值 $\mathbf{x}_i$。设 $\boldsymbol{\beta}_i$ 的共同均值为 $\boldsymbol{\beta}$。此数据生成过程可重新写成：

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + (u_i + \mathbf{x}_i'(\boldsymbol{\beta}_i - \boldsymbol{\beta}))$$

同时做出足够多的假设，以确保回归元 $\mathbf{x}_i$ 与误差项 $(u_i + \mathbf{x}_i'(\boldsymbol{\beta}_i - \boldsymbol{\beta}))$ 是不相关的。因此，y 关于 $\mathbf{x}$ 的 OLS 回归元能够一致地估计出 $\boldsymbol{\beta}$，注意，即使 u_i 是同方差的，误差 $(u_i + \mathbf{x}_i'(\boldsymbol{\beta}_i - \boldsymbol{\beta}))$ 也是异方差的。

对于面板数据而言，标准模型就是随机效应模型（参见 21.7 节），该模型设截距随不同个体而变化，而斜率系数却不是随机的。

对于非线性模型，类似结果不一定成立，而随机参数模型因其允许更丰富的参数结构而受到人们的青睐。当个体对 $\mathbf{x}$ 变化存在异质性响应时，随机参数模型是一致的。一个重要事例是 15.7 节的随机参数 logit。

当个体的回归参数 $\boldsymbol{\beta}_i$ 与可观测个体特征有关时，会产生更严重的复杂性。于是，OLS 估计能导致非一致的参数估计。一个事例就是面板数据的固定效应模型（参见 21.6 节），在该模型中，y 对 $\mathbf{x}$ 的 OLS 估计是非一致的。在此事例中，但不是在所有这样的事例中，存在回归参数子集上的可供选择的一致估计量。

4.8 工具变量

在微观经济计量学中，值得强调的重要复杂情况，是由内生回归元引起的非一致性参数估计的可能性。于是，回归估计便仅仅测算出关联的数值大小，而不是起因的数量及方向；而对于政策分析来说，这两者都是需要的。

然而，工具变量估计量提供了获得一致参数估计的方法。这种方法广泛地用于经济计量学领域，却极少用于其他方面，因为它的概念令人感觉晦涩难懂，并且很可能被误用。

我们将详细加以阐述，先定义工具变量，然后解释工具变量法如何在抽样背景下起作用。

4.8.1 *OLS 的一致性*

考察只有因变量 y 与单一回归元 x 的纯量回归模型。回归分析的目的是估计条件均值函数 $\mathrm{E}[y|x]$。为了记号简洁方便，把没有截距项的线性条件均值模型设定为：

$$\mathrm{E}[y|x] = \beta x \tag{4.42}$$

如果因变量与回归元变量都以它们各自的平均偏差表示，那么没有截距的模型就

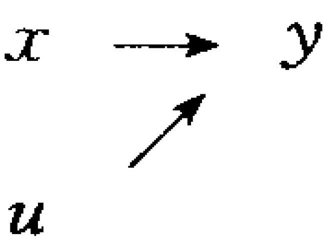
可纳入具有截距的模型之中。关注内容在于获得 β 的一致估计值，因为给定 x 外生变化时，这会提供条件均值的变动。例如，关注内容可以是由归于外生原因的受教育增加而引起的工资效应，譬如增大学生离校的最低年龄，这个决策不是由个体选择决定的。

OLS 回归模型设定为：

$$y=\beta x+u \tag{4.43}$$

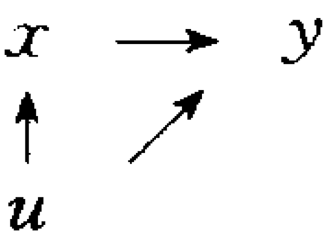
其中，u 表示误差项。y 对 x 进行回归会得到 β 的 OLS 估计值 $\hat{\beta}$。

一些标准回归结果均做出下述假设：模型(4.43)中的回归元与误差项是不相关的。于是，x 对 y 的唯一效应是通过 βx 项而获得的直接效应。我们拥有下面的路径分析图：



其中，x 与 u 之间不存在关联。因而，x 与 u 是 y 的独立原因。

然而，在一些情况下，回归元与误差项之间可能存在关联。例如，考察工资对数(y)对受教育年数(x)的回归。误差项 u 包括了除受教育决定工资之外的所有因素，诸如能力。假定一个人具有很高的 u 值，这是由于(不可观测的)高能力而引起的。因为 $y=\beta x+u$，所以会增加工资。但是，高能力也会导致较大的 x，因为对那些具有高能力的人而言，所受的教育可能也较高。于是，更适宜的路径分析如下：



其中，x 与 u 之间现在存在关联。

x 与 u 之间的这种相关性后果是什么呢？现在，较大的 x 对 y 拥有两个效应。由式(4.43)知，一种直接效应由 βx 而产生，另一种间接作用效应经由 u 而影响到 x，这反过来影响 y。回归的目的只是估计第一种效应，得到 β 的估计值。然而，一旦此事例得出 $\hat{\beta}>\beta$，即两种效应都是正的，OLS 估计将会兼有两种效应。若利用微分计算，我们可得到对 $y=\beta x+u(x)$ 的全微分：

$$\frac{dy}{dx}=\beta+\frac{du}{dx} \tag{4.44}$$

由数据可得到 dx/dy 的信息，因而 OLS 估计出全效应 $\beta+dx/dy$，而不是单独的 β。因此，OLS 估计量是有偏的，且关于 β 为非一致的，除非 x 与 u 之间不存在关联。

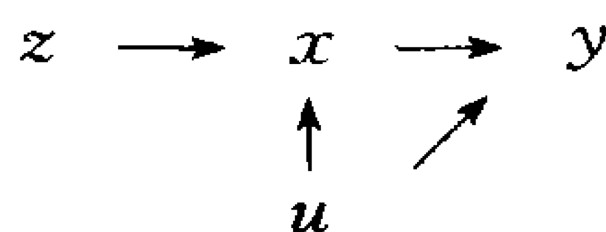
对具有 K 个回归元的线性回归模型进行更正式研究，会得到同样的结论。由 4.7.1 节知，OLS 一致性的必要条件是 $\text{plim}\, N^{-1}\mathbf{X}'\mathbf{u}=\mathbf{0}$。一致性需要回归元在渐近形式上与误差项是不相关的。由式(4.37)知，OLS 非一致性的数量大小是 $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$，即源自 u 对 $\mathbf{x}$ 回归的 OLS 系数。这恰好是 $du/d\mathbf{x}$ 的 OLS 估计值，从而证实了式(4.44)的直观结果。

4.8.2 工具变量

OLS 的非一致性归因于 x 的内生性，意味着 x 的变化不仅与 y 的变化有联系，而且与误差 u 有联系。所需要的内容是可生成 x 唯一外生变化的方法。一种明显的方法是通过随机化实验，但对绝大多数经济应用来说，这类实验成本太高或者甚至行不通。

工具定义

一种原始实验方法或者处理方法利用观测数据仍然是可行的，倘若存在工具（**instrument**）z，z 具有如下性质：z 的变化与 x 的变化有联系，但并不会引起 y 的变化（除了通过 x 的间接途径之外）。这就产生下面这个路径图：



这里，引入变量 z，z 在因果关系上与 x 有关而与 u 无关。还可以是下述情况：z 与 y 是相关的，但这种相关性的唯一来源是，z 与 x 成为相关的间接途径，这反过来决定 y。z 作为 y 模型中回归元的更为直接的途径被排除。

如果：(1) z 与误差项 u 无关；(2) z 与回归元 x 相关，更正式地，变量 z 称为纯量回归模型 $y=\beta x+u$ 中关于回归元 x 的工具（**instrument**）或工具变量（**instrument variables**）。

第一个假设排除工具 z 成为关于 y 模型的回归元，然而，如果 y 既依赖于 x 又依赖于 z，而且 y 只对 x 进行回归，那么 z 被并入到误差之中，从而 z 与 u 也就相关了。第二个假设要求工具与作为工具的变量之间存在某种关联。

工具的事例

在许多微观经济计量应用中，很难找到合理的工具。此处，我们给出两个事例。

首先，假设我们要估计外生的市场价格变化所引起的市场需求响应。显然，需求量依赖于价格，但价格不是外生的，因为已知它们是部分地由市场需求来决定的。一个合适的价格工具是这样的变量，该变量与价格相关，但又并不直接影响需求量。一个明显的备选者就是影响市场供给的变量，因为这一变量也影响价格，但不直接决定需求。如果对农产品进行建模，那么一个事例就是对有利栽培条件的测量。倘若有利栽培的条件不直接影响需求，并且大大得益于正式的供给与需求的经济模型，则此种工具的选择是无争议的。

其次，假设我们要估计受教育外生变化而引起的收益。绝大多数观测数据集合均缺少对个体能力的测量，因而工资对受教育的回归包含不可观测能力的误差，进而与受教育回归元相关。我们需要找到一个工具 z，使其与受教育相关且与能力无关，并且更一般地，与误差项无关，这意味着工具不能直接决定工资。

关于 z 的一种流行的备选者是接近于学院或者大学的程度[卡德（Card，1995）]。显然，这满足第 2 个条件，例如，距离社区学院或者州立大学较远的人不太可能上大学。它很可能满足第 1 个条件，虽然可以证明，那些住所距离学院很远

的人可能在低工资劳动力市场中，人们需要估计 y 的多元回归，这包括另外一些回归元，比如代表非大都市区域的标示变量。

工具的第二个备选者是出生月份[安格里斯特和克鲁格(Angrist and Krueger，1991)]。显然，这一工具满足第 1 个条件，因为没有理由认为，若回归中包括年龄的话，出生月份对工资拥有直接影响。令人惊讶的是，第 2 个条件也可能得到满足，因为在美国，出生月份决定最初入学年龄，一些法律规定最小离校年龄，这反过来可能影响到受教育年数。邦德、耶格和贝克(Bound, Jaeger and Baker, 1995)对此工具曾给出评论。

4.9 节将详细讨论选取不适当工具的结果。

4.8.3 工具变量估计量

对于具有纯量回归元 x 与纯量工具 z 的回归，工具变量估计量[**instrumental variables (IV) estimator**]被定义为：

$$\hat{\beta}_{IV}=(\mathbf{z}'\mathbf{x})^{-1}\mathbf{z}'\mathbf{y} \tag{4.45}$$

在纯量回归元的情况下，$\mathbf{z}$、$\mathbf{x}$、$\mathbf{y}$ 均表示 $N\times 1$ 维向量。如果 z 与 x 相关，而与误差项不相关，那么这个估计量就给出线性模型 $y=\beta x+u$ 斜率系数 β 的一致估计。

存在几种方法推导式(4.45)。我们提供一种直观的推导，该方法不同于通常的譬如 6.2.5 节阐述的推导。

回到受教育—工资的事例上。假定工具 z 变动 1 个单位，与之关联的受教育就多 0.2 年，而且年工资会增加 500 美元。这种工资上的增加是 z 增大导致受教育年数增加的间接影响的后果，这反过来促使收入增多。于是，由此可见，受教育多增加 0.2 年，就会使工资增加 500 美元，因此，受教育多增加 1 年，会使工资增加 2 500 美元(500/0.2)。因而，β 的原因估计值是 2 500。利用数学记号表示，我们估计 dx/dz 与 dy/dz 的变化，并计算出原因估计量如下：

$$\beta_{IV}=\frac{dy/dz}{dx/dz} \tag{4.46}$$

这种证明原因参数 β 的方法是由赫克曼(Heckman, 2000，第 58 页)给出的；也可参见 2.4.2 节的事例。

剩下的内容是对 dx/dz 与 dy/dz 进行一致估计。估计 dy/dz 的一种显而易见的方法，是通过 y 对 z 的 OLS 回归进行估计，其斜率估计值为$(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{y}$。类似地，通过 x 对 z 的 OLS 回归，可以估计 dy/dz，其斜率估计值为$(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{x}$。于是：

$$\hat{\beta}_{IV}=\frac{(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{y}}{(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{x}}=(\mathbf{z}'\mathbf{x})^{-1}\mathbf{z}'\mathbf{y} \tag{4.47}$$

4.8.4 沃尔德估计量

IV 的一个重要而简单的事例，是工具 z 作为一个二值工具(**binary instrument**)。当 $z=1$ 时，分别用 $\bar{y}_1$ 与 $\bar{x}_1$ 表示 y 与 x 的子样本均值；而当 $z=0$ 时，分别用 $\bar{y}_0$ 与 $\bar{x}_0$ 表示 y 与 x 的子样本均值。于是有 $\Delta y/\Delta z=(\bar{y}_1-\bar{y}_0)$和 $\Delta x/\Delta z=(\bar{x}_1-\bar{x}_0)$，从

而由式(4.46)得到：

$$\hat{\beta}_{\text{Wald}}=\frac{\bar{y}_1-\bar{y}_0}{\bar{x}_1-\bar{x}_0} \tag{4.48}$$

此估计量被命名为沃尔德估计量(**Wald estimator**)，或者称为分组估计量(**grouping estimator**)。

沃尔德估计量还能从公式(4.45)中获得。对于没有截距的模型来说，变量是以偏离其均值多少而测量的，因而 $\mathbf{z}'\mathbf{y}=\sum_i(z_i-\bar{z})(y_i-\bar{y})$。就二值工具 z 而言，这会得出 $\mathbf{z}'\mathbf{y}=N_1(\bar{y}_1-\bar{y})=N_1N_0(\bar{y}_1-\bar{y}_0)/N$，其中，$N_0$ 与 N_1 分别表示那些对应 $z=0$ 与 $z=1$ 的观测值个数。该结果使用了 $\bar{y}_1-\bar{y}=(N_0\bar{y}_1+N_1\bar{y}_1)/N-(N_0\bar{y}_0+N_1\bar{y}_1)/N=N_0(\bar{y}_1-\bar{y}_0)/N$。类似地，$\mathbf{z}'\mathbf{x}=N_1N_0(\bar{x}_1-\bar{x}_0)/N$。结合这些结果，我们可通过式(4.45)推导出式(4.48)。

对于受教育工资事例，已经假定我们能定义出两个组，每一组的隶属关系并不能直接决定工资，尽管它会影响到受教育水平，进而间接影响到工资。然后，IV 估计值是两个组的平均工资之差被两个组的平均受教育之差去除。

4.8.5 *样本协方差与相关性分析*

IV 估计量还可以用协方差或相关性给出解释。

对于样本协方差，我们直接从式(4.45)得到：

$$\hat{\beta}_{\text{IV}}=\frac{\text{Cov}[z,y]}{\text{Cov}[z,x]} \tag{4.49}$$

上式中，用 Cov[ ]表示样本协方差。

对于样本相关性，注意到，模型(4.43)的 OLS 估计量能写成 $\hat{\beta}_{\text{OLS}}=r_{xy}\sqrt{\mathbf{y}'\mathbf{y}}/\sqrt{\mathbf{x}'\mathbf{x}}$，其中，$r_{xy}=\mathbf{x}'\mathbf{y}/\sqrt{(\mathbf{x}'\mathbf{x})(\mathbf{y}'\mathbf{y})}$表示 x 与 y 的样本相关(**sample correlation**)系数。这就导致把 OLS 估计量解释为，x 变动一个标准差而引起 y 变动了 r_{xy} 标准差。问题是，相关系数 r_{xy} 被 x 与 u 之间的相关性关系所混淆。一种可供选择的方法是，通过用 z 与 y 之间的相关系数除以 z 与 x 之间的相关系数，来间接地测算 x 与 y 的相关系数。于是有：

$$\hat{\beta}_{\text{IV}}=\frac{r_{zy}\sqrt{\mathbf{y}'\mathbf{y}}}{r_{zx}\sqrt{\mathbf{x}'\mathbf{x}}} \tag{4.50}$$

可以证明，它等于式(4.45)中的 $\hat{\beta}_{\text{IV}}$。

4.8.6 *多元回归的 IV 估计*

现在，考察具有特殊观测值

$$y=\mathbf{x}'\boldsymbol{\beta}+u$$

的多元回归模型，它拥有 K 个回归元变量，因而 $\mathbf{x}$ 与 $\boldsymbol{\beta}$都是 $K\times 1$ 维向量。

工具

假设存在 $r\times 1$ 维工具向量 $\mathbf{z}$，且 $r\geqslant k$，满足如下条件：

1. $\mathbf{z}$ 与误差 u 不相关；

2. $\mathbf{z}$ 与回归向量 $\mathbf{x}$ 是相关的；

3. $\mathbf{z}$ 与回归元向量 $\mathbf{x}$ 是强相关的，而不是弱相关的。

对于一致性而言，前两个性质是必需的，并且前面已对纯量情况进行了阐述。定义于 4.9.1 节的第三个性质是为了确保 IV 估计量具有良好的有限样本性能而对第二个性质的强化。

在多元回归情况下，$\mathbf{z}$ 与 $\mathbf{x}$ 可能分享某些共同的分量。$\mathbf{x}$ 的某些分量可能与 u 是不相关的，这些分量称为外生回归元（**exogenous regressors**）。显然，这些分量因其满足第 1 个条件与第 2 个条件而适合作为工具。$\mathbf{x}$ 的另外一些分量可能与 u 是相关的，这样的分量称为内生回归元（**endogenous regressors**）。这些分量会导致 OLS 的非一致性，并且明显不适合作为工具，因为它们满足第 1 个条件。把 $\mathbf{x}$ 分割成$\mathbf{x}=[\mathbf{x}_1' \ \ \mathbf{x}_2']$，其中，$\mathbf{x}_1$ 包含内生回归元，而 $\mathbf{x}_2$ 包含外生回归元。于是，有效的工具是 $\mathbf{z}=[\mathbf{z}_1' \ \ \mathbf{x}_2']$，其中，$\mathbf{x}_2$ 能作为自身的工具，但是，我们需要找到至少与已有内生变量 $\mathbf{x}_1$ 一样多的工具 $\mathbf{z}_1$。

识别

联立方程模型的识别已在 2.5 节阐述。这里，我们具有单个方程。阶条件（**order condition**）要求工具的个数至少等于独立内生分量的个数，所以 $r \geqslant K$。若 $r=K$，则此模型称为恰好识别的（**just-identified**）；若 $r>K$，则此模型称为过度识别的（**overidentified**）。

在大量多元回归应用中，仅有一个内生回归元。例如，工资对受教育回归将包括很多其他回归，譬如年龄、地理位置以及家庭背景。关注内容为受教育的系数，但这是最可能与误差项相关的是内生变量，因为能力是不可观测的。关于受教育的必需的单个工具的可能备选者已由 4.8.2 节给出。

如果工具不满足第 1 个条件，那么该工具就是无效工具（**invalid instrument**）。如果工具不满足第 2 个条件，那么该工具就是不相关工具[1]（**irrelevant instrument**）。如果极少的工具是相关的，那么此模型可能是不可识别的（**unidentified**）。当工具与作为工具的那个内生变量之间存在很小的相关性时，第 3 个条件就不成立。此模型称为弱识别的（**weakly identified**），而该工具称为弱工具（**weak instrument**）。

工具变量估计量

当模型是恰好识别时，有 $r=K$，工具变量估计量显然就是对式（4.45）中矩阵的推广：

$$\hat{\beta}_{\mathrm{IV}}=(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y} \tag{4.51}$$

其中，$\mathbf{Z}$ 表示 $N\times K$ 阶矩阵，第 i 行是 $\mathbf{z}_i'$。一旦把式（4.51）中的 $\mathbf{y}$ 用回归模型 $\mathbf{y}=\mathbf{X}\beta+\mathbf{u}$ 代入，得到：

---

〔1〕 又称为不相干工具。——译者注

$$
\begin{aligned}
\hat{\beta}_{\mathrm{IV}} &= (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'[\mathbf{X}\beta+\mathbf{u}] \\
&= \beta+(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{u} \\
&= \beta+(N^{-1}\mathbf{Z}'\mathbf{X})^{-1}N^{-1}\mathbf{Z}'\mathbf{u}
\end{aligned}
$$

由此可得，若：

$$\operatorname{plim} N^{-1}\mathbf{Z}'\mathbf{u}=\mathbf{0}$$

且

$$\operatorname{plim} N^{-1}\mathbf{Z}'\mathbf{X}\neq\mathbf{0}$$

则 IV 估计量是一致的。这些条件本质上就是第 1 个条件与第 2 个条件，即 $\mathbf{z}$ 与 $\mathbf{u}$ 是不相关的，而 $\mathbf{z}$ 与 $\mathbf{x}$ 是相关的。为了确保 $N^{-1}\mathbf{Z}'\mathbf{X}$ 的逆存在，假定 $\mathbf{Z}'\mathbf{X}$ 是满秩的，其秩为 K。这是比阶条件 $r=K$ 稍强的假设。

就异方差误差而言，IV 估计量在渐近形式上是正态的，其均值为 β，而且方差矩阵可通过：

$$\hat{\mathrm{V}}[\hat{\beta}_{\mathrm{IV}}]=(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\hat{\mathbf{\Omega}}\mathbf{Z}(\mathbf{X}'\mathbf{Z})^{-1} \tag{4.52}$$

一致地估计出来，其中，$\hat{\mathbf{\Omega}}=\mathrm{Diag}[\hat{u}_i^2]$。此结果可利用类似于 4.4.4 节给出 OLS 的方式求出。

虽然 IV 估计量是一致的，但它在实践上却损失了相当大的有效性。直观上讲，如果工具 $\mathbf{z}$ 与回归元 $\mathbf{x}$ 具有很小的相关性，那么 IV 将不会起作用(参见 4.9.3 节)。

4.8.7 *两阶段最小二乘法*

式(4.51)中的 IV 估计量要求工具的数量与回归元的数量相等。对于过度识别模型来说，可通过去掉一些工具而使该模型变为恰好识别的，这样就可利用 IV 估计量。但是，当去掉一些工具变量时，会发生渐近有效性损失。

然而，一种普遍方法是使用两阶段最小二乘法(2SLS)：

$$\hat{\beta}_{\mathrm{2SLS}}=[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}] \tag{4.53}$$

此方法将在 6.4 节阐述，并且解释其动机。

2SLS 估计量是 IV 估计量。在恰好识别模型中，它简化成由式(4.51)给出的含有工具 $\mathbf{Z}$ 的 IV 估计量。在过度识别模型中，如果工具是 $\hat{\mathbf{X}}$，那么 2SLS 估计量等于式(4.51)给出的 IV 估计量，其中，$\hat{\mathbf{X}}=\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ 表示 $\mathbf{x}$ 对 $\mathbf{z}$ 的 OLS 回归预测值。

顾名思义，2SLS 估计量是通过两次连续关联的 OLS 回归而获得的结果：先通过 $\mathbf{x}$ 对 $\mathbf{z}$ 的 OLS 回归得到 $\hat{\mathbf{x}}$，再通过 y 对 $\hat{\mathbf{x}}$ 的 OLS 回归得到 $\hat{\beta}_{\mathrm{2SLS}}$。这种解释不一定能推广到非线性模型，参见 6.5.4 节。

2SLS 估计量经常以更紧凑的形式表述成：

$$\hat{\beta}_{\mathrm{2SLS}}=[\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X}]^{-1}[\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{y}] \tag{4.54}$$

其中：

$$\mathbf{P}_{\mathbf{Z}}=\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$$

表示幂等投影矩阵(**projection matrix**)，它满足 $\mathbf{P}_{\mathbf{Z}}=\mathbf{P}_{\mathbf{Z}}'$、$\mathbf{P}_{\mathbf{Z}}\mathbf{P}_{\mathbf{Z}}'=\mathbf{P}_{\mathbf{Z}}$，以及 $\mathbf{P}_{\mathbf{Z}}\mathbf{Z}=\mathbf{Z}$。

可以证明,2SLS 估计量是渐近正态的,其估计渐近方差为:

$$\hat{\mathrm{V}}[\hat{\boldsymbol{\beta}}_{2SLS}]=N[\mathbf{X}'\mathbf{P_Z}\mathbf{X}]^{-1}[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\hat{\mathbf{S}}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}][\mathbf{X}'\mathbf{P_Z}\mathbf{X}]^{-1} \tag{4.55}$$

在通常的异方差误差情况下,$\hat{\mathbf{S}}=N^{-1}\sum_i \hat{u}_i^2\mathbf{z}_i\mathbf{z}_i'$ 与 $\hat{u}_i=y_i-\mathbf{x}_i'\hat{\boldsymbol{\beta}}_{2SLS}$。广泛使用的小样本调整,是在 $\hat{\mathbf{S}}$ 公式中用 $N-K$ 去除,而不用 N 去除。

在误差项是同方差的特殊情况下,可进行简化,并且 $\hat{\mathrm{V}}[\hat{\boldsymbol{\beta}}_{2SLS}]=s^2[\mathbf{X}'\mathbf{P_Z}\mathbf{X}]^{-1}$。在许多引论课程的处理中,都给出了后面的结果,但是,更一般的公式(4.55)倾向于现代方法,即把误差项当成潜在的异方差项。

对于具有异方差误差的过度识别模型而言,被怀特(White, 1982)称为两阶段工具变量估计量(**two-stage instrumental variables estimator**)的那种估计量,比 2SLS 更为有效。此外,一些广泛使用的模型设定检验需要通过这一估计量而不是 2SLS 加以估计。有关详细内容,参见 6.4.2 节。

4.8.8 *IV* 的范例

举一个 IV 估计的事例,考察数据生成过程为:

$$y=0+0.5x+u$$
$$x=0+z+v$$

对 x 的斜率系数进行估计,其中 $z\sim\mathcal{N}[2,1]$,并且(u,v)为联合正态分布,其均值为 0,方差为 1,相关系数为 0.8。

y 对 x 的 OLS 估计是非一致的,这是因为由构造知,x 与 v 相关,所以 x 与 u 相关。IV 估计会得出一致估计值。由结构知,z 与 u 是不相关的,但与 x 是相关的,所以变量 z 是有效工具。z 的一些变形,比如 z^3,也都是有效工具。

表 4.4 给出源自样本量为 10 000 的生成数据的各种估计值与相关的标准误差。我们关注斜率系数。

表 4.4　工具变量例子[a]

	OLS	IV	2SLS	IV (z^3)
常值	−0.804	−0.017	−0.017	−0.014
	(0.014)	(0.022)	(0.032)	(0.025)
x	0.902	0.510	0.510	0.509
	(0.006)	(0.010)	(0.014)	(0.012)
R^2	0.709	0.576	0.576	0.574

[a] 样本量为 10 000 的生成数据。OLS 是非一致的,其他三个估计量是一致的。当误差是同方差时,报告出了稳健标准误差,不过这里没有将它们写出来。2SLS 标准误差是不合适的。数据生成过程由下一节给出。

OLS 估计量是非一致的,其斜率系数估计值 0.902 比出自真实值 0.5 的 50 个标准差还要大一些。其余的估计值是一致的,并且都位于 0.5 的两个标准差之内。

存在几种方法计算 IV 估计量。出自 y 对 z 的 OLS 回归所得到的斜率系数是 0.516 8，而出自 x 对 z 的 OLS 回归所得到的斜率系数是 1.012 4，从而利用式(4.47)，得到 IV 估计值为 0.516 8/1.012 4=0.510。在实践中，人们把 z 作为 x 的工具，同时利用式(4.52)计算标准误差，而不是利用式(4.45)或式(4.51)直接计算 IV 估计量。2SLS 估计量[参见式(4.54)]能通过 y 对 $\hat{x}$ 的 OLS 回归来进行计算，其中，$\hat{x}$ 表示 x 对 z 的 OLS 回归的预测值。在这一恰好识别的模型中，虽然如同在 6.5.4 节将要解释的，出自 y 对 $\hat{x}$ 的 OLS 回归的标准差是错误的，但 2SLS 估计值精确地等于 IV 估计值。

最后一列用 z^3 而不是 z 作为 x 的工具。因为 z^3 与 u 不相关，而与 x 相关，所以这个可供选择的 IV 估计量是一致的。然而，对此特定 dgp 而言，z^3 表现出缺乏有效性，同时斜率系数的标准误差由 0.010 上升到 0.012。

对于单个回归元与单个工具情况下的一般结果来说，与 OLS 估计相比，IV 估计的有效性存在损失，参见式(4.61)。此处，$r_{x,z}^2=0.510$ 是较大的，这在表 4.4 中没有给出，因而其损失并不大，但斜率系数的标准误差却从 0.006 略微增大到 0.010。在实践中，有效性的损失比这要大得多。

4.9 实践中的工具变量

重要的实践问题包括，决定 IV 方法是否是必需的，如果是必需的，则决定工具是否是有效的。有关的设定检验将在 8.4 节中阐述。不幸的是，检验的有效性是有局限性的。检验需要下述假设：恰好识别模型中的工具是有效的，同时只对过度识别约束进行检验。

给定有效工具时，虽然 IV 估计量是一致的，正如下面将要阐述的，但是同 OLS 估计量相比，IV 估计量的有效性大打折扣且拥有有限样本分布，同时，通常的有限样本量与渐近分布的情况截然不同。如果工具与要作为工具的那个变量之间存在弱相关性，那么这个问题将被放大。若存在的工具比所需要的更多一些，则会出现弱工具。这可直接通过去掉一些工具来加以处理[参见唐纳德和纽韦(Donald and Newey，2001)]。甚至当具有最少工具个数，并且有一个或多个工具是弱工具时，就会产生更为基础性的问题。

4.9.1 *弱工具*

弱工具不存在单独一种定义。许多学者使用如下的弱工具(**weak instrument**)标志，此处以逐渐增加复杂模型的形式加以阐述。

- 纯量回归元 x 与纯量工具 z：弱工具是指，使 $r_{x,z}^2$ 很小的工具。
- 纯量回归元 x 与向量工具 $\mathbf{z}$：如果源自 x 对 $\mathbf{z}$ 的回归所得到的 R^2(记为 $R_{x,\mathbf{z}}^2$)很小，或者该回归整体拟合的 F 统计量很小，那么此工具就是弱工具。
- 多元回归元 $\mathbf{x}$ 只含有一个内生的分量：弱工具是指使偏 R^2(**partial $\mathbf{R}^2$**)很小或者 F 统计量很小的工具，这些偏统计量将在 4.9.1 节末尾加以定义。
- 多元回归元 $\mathbf{x}$ 含有几个内生分量：存在几种测量方法。

R^2 测量

考虑单方程：

$$y=\beta_1 x_1+\mathbf{x}_2'\boldsymbol{\beta}_2+u \tag{4.56}$$

其中，只有一个回归元 x_1 是内生的，而其余回归元向量 $\mathbf{x}_2$ 是外生的。假定工具向量 $\mathbf{z}$ 包含外生变量 $\mathbf{x}_2$ 和至少一个其他的工具。

一个可行 R^2 测量是源自 x_1 对 $\mathbf{z}$ 回归的通常的 R^2。然而，因为 x_1 与 $\mathbf{x}_2$ 是高度相关的，所以 R^2 可能会很高，但是，从直观上看，我们确实需要 x_1 与工具而不是与 $\mathbf{x}_2$ 高度相关。

因此，邦德、耶格和贝克(Bound, Jaeger and Baker, 1995)提出使用偏 $\mathbf{R}^2$，记为 R_p^2，它能清除 $\mathbf{x}_2$ 的影响。R_p^2 可从回归：

$$(x_1-\tilde{x}_1)=(\mathbf{z}-\tilde{\mathbf{z}})\gamma+v \tag{4.57}$$

中获得 R^2 作为 R_p^2，其中，$\tilde{x}_1$ 与 $\tilde{\mathbf{z}}$ 分别表示 x_1 对 $\mathbf{x}_2$ 回归与 $\mathbf{z}$ 对 $\mathbf{x}_2$ 回归的拟合值。在恰好识别的情况下，$\mathbf{z}-\tilde{\mathbf{z}}$ 将简化成 $z_1-\tilde{z}_1$，其中，z_1 表示单个工具而不是 $\mathbf{x}_2$，而 $\tilde{z}_1$ 表示源自 z_1 对 $\mathbf{x}_2$ 回归的拟合值。

R_p^2 比 $R_{x_1,\mathbf{z}}^2$ 要小许多，这并不令人感到意外。当只存在唯一回归元并且它是内生的时，R_p^2 的公式可以简化成 r_{xz}^2，而当仅存在一个工具时，它进一步简化成 $\mathrm{Cor}[x, z]$。

考察具有一个以上内生变量的单方程模型，同时关注第一个内生变量的系数估计。于是，式(4.56)中，x_1 是内生的，另外，$\mathbf{x}_2$ 中的一些变量也是内生的。几种可供选择的其他测量方式是用控制其他内生回归元存在的残差替代式(4.57)的右边项。谢伊(Shea, 1997)提出一种偏 R^2，比如说 R_p^{*2}，把它计算成$(x_1-\tilde{x}_1)$与$(\hat{x}_1-\tilde{\hat{x}}_1)$之间样本相关的平方，这里的$(x_1-\tilde{x}_1)$再次表示源自 x_1 对 $\mathbf{x}_2$ 回归的残差，而$(\hat{x}_1-\tilde{\hat{x}}_1)$表示源自$\hat{x}_1$(来自 x_1 对 $\mathbf{z}$ 回归的拟合值)对$\hat{x}_2$(来自 $\mathbf{x}_2$ 对 $\mathbf{z}$ 回归的拟合值)回归的残差。这里，$\hat{x}_1$ 是 x_1 关于 $\mathbf{z}$ 回归的拟合，而 $\hat{\mathbf{x}}_2$ 是 $\mathbf{x}_2$ 关于 $\mathbf{z}$ 的拟合。波斯基特和斯基尔斯(Poskitt and Skeels, 2002)提供一种可供选择的偏 R^2，它如同谢伊的 R_p^{*2} 一样，当只存在一个内生回归元时，就简化为 R_p^2。然而，霍尔、鲁德布施和威尔克科斯(Hall, Rudebusch and Wilcox, 1996)提出使用典型相关。

这些关于第一个内生变量系数的测量法，也可对其他的内生变量重复进行。另外，波斯基特和斯基尔斯(Poskitt and Skeels, 2002)考察了可应用于所有内生变量联合工具的 R^2 测量。

当偏 R^2 测量失效时，估计量的非一致性问题与预测损失将被放大，正如在 4.9.2 节和 4.9.3 节详细阐述的。特别地，参见式(4.60)和式(4.62)。

偏 F 统计量

对于将在 4.9.4 节考察的不好的有限样本表现，普遍使用相对测量，即使用在内生回归元对工具的回归中系数是否为 0 的 F 统计量。

对于内生的单个回归元，我们使用整体 F 统计量，检验内生回归元对工具的回归 $x=\mathbf{z}'\boldsymbol{\pi}+v$ 中，$\boldsymbol{\pi}=\mathbf{0}$。

更广泛地，一些外生回归元也可以出现在模型中，而在具有单个内生回归元 $\mathbf{x}_1$

的模型(4.56)中,我们运用:

$$x = \mathbf{z}_1' \boldsymbol{\pi}_1 + \mathbf{x}_2' \boldsymbol{\pi}_2 + v \tag{4.58}$$

中 $\boldsymbol{\pi}_1 = \mathbf{0}$ 检验的 F 统计量,其中,$\mathbf{z}_1$ 表示工具而不是外生回归元,$\mathbf{x}_2$ 表示外生回归元。这是两阶段最小二乘法 IV 解释的第一阶段回归。

此统计量用作 IV 估计量中潜在有限样本偏倚的标志。在 4.9.4 节,我们将解释施泰格和斯托克(Staiger and Stock, 1997)的结果,他们提出,小于 10 的值是有问题的,而小于或等于 5 的值则是极端有限样本偏倚的标志,同时,我们考虑把它推广到一个以上内生回归元的情况。

4.9.2 *IV 估计量的非一致性*

IV 具有一致性的本质条件是 4.8.6 节中的第 1 个条件,即工具应该与误差项不相关。在恰好识别情况下,不存在这类检验。在过度识别情况下,对过度识别假设进行检验是可能的(参见 6.4.3 节)。于是,拒绝可能是由于工具的内生性或由于模型失效所导致。因此,第 1 个条件很难被直接检验,并且决定一个工具是否为外生的,通常是很主观的决策,尽管人们经常以经济理论为指导。

通过函数形式约束(**function form restriction**)创立外生工具总是可行的。例如,假定存在两个回归元,因而 $y=\beta_1 x_1+\beta_2 x_2+u$,其中,$x_1$ 与 u 是不相关的,x_2 与 u 是相关的。注意到,本节自始至终地假定所有变量均是以偏离其均值来测算的,所以,为了不失一般性,省略截距项。于是,OLS 是非一致的,因为 x_2 是内生的。看起来,关于 x_2 的好工具是 x_1^2,因为 x_1 与 u 是不相关的,所以 x_1^2 与 u 也是不相关的。然而,这个工具的有效性需要有关条件均值的函数形式约束,即 x_1 仅以线性方式而不是以二次形式进入模型。在实践中,人们认为,线性模型是唯一的近似,并且以这种人工方式获得的工具很容易遭到批评。

创立有效工具的一种较好的方式,是通过可供选择的排除性约束(**exclusion restrictions**),该排除性约束并非十分依赖于对函数形式的选择。几个实践例子由 4.8.2 节给出。

一些结构模型,比如经典线性联立方程模型(参见 2.4 节和 6.10.6 节),均以非常明显的方式做出这类排除性约束。由于约束太有针对性而常常遭到批评,除非有令人信服的经济理论支持这些约束。

对于面板数据应用来说,有理由假定当前仅有的数据可用于关注的方程之中——在误差项是序列不相关的假设下,排除性约束允许把过去数据用作工具(参见 22.2.4 节)。类似地,在不确定性下进行决策的模型中(参见 6.2.7 节),把滞后变量用作工具,因为它们是信息集的一部分。

不存在工具外生性的正式检验,工具外生性没有另外去检验回归方程是否被正确设定。工具外生性必然依赖先验信息,比如来自经济学理论或统计理论的信息。邦德等人(Bound et al., 1995,第 446~447 页)对安格里斯特和克鲁格(Angrist and Krueger, 1991)所使用工具有效性的评价,提供了涉及确定工具外生性的杰出的事例。

特别重要的是，如果工具是弱的，那么工具就是外生的，因为就弱工具而言，甚至工具的非常适度的内生性能导致 IV 参数估计值，该值与已经非一致的 OLS 参数估计值相比，显得更加非一致。

简单起见，考察具有一个回归元与一个工具的线性回归模型，因此，$y=\beta x+u$。然后，经过一些代数计算（留作习题），得出：

$$\frac{\text{plim}\,\hat{\beta}_{\text{IV}}-\beta}{\text{plim}\,\hat{\beta}_{\text{OLS}}-\beta}=\frac{\text{Cor}[z,u]}{\text{Cor}[x,u]}\times\frac{1}{\text{Cor}[z,x]} \tag{4.59}$$

因而，就无效工具以及工具与回归元之间很小的相关性而言，IV 估计量甚至比 OLS 的非一致性更加严重。例如，假定 z 与 x 的相关系数是 0.1，这对横截面数据来说也是不罕见的。于是，只要 z 与 u 的相关系数大于 x 与 u 的相关系数 0.1 倍，IV 的非一致性比 OLS 的非一致性更严重。

结果(4.59)可以推广到具有一个内生回归元和几个外生回归元、具有 i. i. d 误差以及工具包含所有外生回归元的模型(4.56)上。从而有：

$$\frac{\text{plim}\,\hat{\beta}_{1,\text{2SLS}}-\beta_1}{\text{plim}\,\hat{\beta}_{1,\text{OLS}}-\beta_1}=\frac{\text{Cor}[\hat{x},u]}{\text{Cor}[x,u]}\times\frac{1}{R_p^2} \tag{4.60}$$

其中，R_p^2 已在式(4.56)后定义。对于一个以上内生回归元情况的推广，参见谢伊(Shea, 1997)。

这些结果对于运用 IV 具有深远的意义，邦德等人(Bound et al., 1995)曾对此加以强调。如果工具是弱的，那么甚至适度的工具内生性，能够导致比 OLS 更为严重的非一致性。也许是因为该结论如此消极，文献忽略了弱工具的这一方面。最近一个著名的例外是哈恩和豪斯曼(Hahn and Hausman, 2003a)。

绝大多数文献均假定第 1 个条件得到满足，所以 IV 是一致性的，同时关注归属于弱工具的其他复杂性。

4.9.3 *低准确性*

当 OLS 是非一致性时，虽然 IV 估计能够产生一致估计，但它还是损失了准确性。从直观上看，由 4.8.2 节知，工具 z 是能够导致 x 上的外生性运动的一个处理，但是这样做会有相当大的噪音。

就弱工具而言，准确性损失会增大，而标准误差则会增加。

随着精度损失的增加，较弱工具的标准差也就增加。在最简单的仅含一个单个回归元与具有 i. i. d 误差项的单一工具情况下，很容易看出这一点。于是，渐近方差为：

$$\begin{aligned}\text{V}[\hat{\beta}_{\text{IV}}]&=\sigma^2(\mathbf{x}'\mathbf{z})^{-1}\,\mathbf{z}'\mathbf{z}(\mathbf{z}'\mathbf{x})^{-1}\\&=[\sigma^2/\mathbf{x}'\mathbf{x}]/[(\mathbf{z}'\mathbf{x})^2/(\mathbf{z}'\mathbf{z})(\mathbf{x}'\mathbf{x})]\\&=\text{V}[\hat{\beta}_{\text{OLS}}]/r_{xz}^2\end{aligned} \tag{4.61}$$

例如，如果 z 与 x 之间的样本相关系数平方等于 0.1，那么 IV 的标准差将是 OLS 的标准差的 10 倍。再者，IV 估计量拥有比 OLS 估计量更大的方差，除非 $\text{Cor}[z,x]=1$。

结果(4.61)能够被推广到具有一个内生回归元以及几个外生回归元、具有 iid 误差以及包括所有外生回归元的工具的模型(4.56)上。从而有：

$$\mathrm{se}[\hat{\beta}_{1,2SLS}]=\mathrm{se}[\hat{\beta}_{1,OLS}]/R_p \tag{4.62}$$

其中，se[·]表示渐近标准差，而 R_p^2 已在式(4.56)后面定义。对于推广到一个以上内生回归元的情况来说，R_p^2 要用谢伊(Shea, 1997)提出的 R_p^{*2} 代替。这就提供了谢伊检验统计量的动机。

差的精确度集中在内生变量的系数上。对于外生变量而言，2SLS 系数估计值的标准误差类似于 OLS 的那些标准误差。从直观上看，外生变量作为其自身的工具，所以，它们确实拥有强工具。

对于内生回归元系数而言，它具有很小的偏 R^2，而不是 R^2，这就导致了估计量精确度的损失。这就解释了 2SLS 标准误差为什么非常大于 OLS 的标准误差，尽管内生变量与工具之间的相关性很高。若利用其他方法，内生变量系数的 2SLS 标准误差远大于 OLS 标准误差，这提供了工具是弱的显著信号。

用于测算由弱工具引起的 IV 精确度损失的统计量称为工具相关(**instrument relevance**)测量法。在某种程度上，如果 IV 标准误差远大于 OLS 标准误差，那么当问题很容易被检测出时，工具相关的测量法就不必要了。

4.9.4 *有限样本偏倚*

本节概述关于“弱工具”的相当富有挑战性的且尚未完成的文献，关注于实践问题，甚至在“大”样本渐近理论下，对于 IV 估计量分布能提供不好的近似。特别地，即使 IV 估计量是渐近一致的，但是在有限样本中，IV 估计量却是有偏的。当工具是弱的时候，此种偏倚尤其显著。

对于非一致性的 OLS 估计量而言，IV 的偏倚可以是相当大的——正如纳尔逊和施塔茨(Nelson and Startz, 1990)通过简单蒙特卡罗实验所证明的，而且邦德等人(Bound et al., 1995)证明，它借助于涉及成千上万个观测值的真实数据，却是非常弱的工具。再者，纳尔逊和施塔茨(Nelson and Startz, 1990)还证明，标准误差也是具有很大偏倚的。

理论文献涉及非常专业且高级的经济计量理论，这是因为获得 IV 估计量样本均值确实很困难。为了理解这一点，考虑适应于由 4.4.8 节给出的 OLS 估计量通常无偏性证明的 IV 估计量。在恰好识别情况下，由式(4.51)定义的 $\hat{\beta}_{IV}$ 得出：

$$\begin{aligned}\mathrm{E}[\hat{\boldsymbol{\beta}}_{IV}]&=\boldsymbol{\beta}+\mathrm{E}_{\mathbf{Z},\mathbf{X},\mathbf{u}}[(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{u}]\\&=\boldsymbol{\beta}+\mathrm{E}_{\mathbf{Z},\mathbf{X}}[(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\times\mathrm{E}[\mathbf{u}|\mathbf{Z},\mathbf{X}]]\end{aligned}$$

其中，利用期望迭代定理(参见 A.8 节)，对于所有随机变量 **Z**、**X** 和 **u**，可通过先对以 **Z** 与 **X** 为条件的 **u** 取数学期望而得到非条件期望值。IV 估计量拥有均值 $\boldsymbol{\beta}$ 的明显充分条件是，$\mathrm{E}[\mathbf{u}|\mathbf{Z},\mathbf{X}]=\mathbf{0}$。然而，这一假设太强，因为它蕴含 $\mathrm{E}[\mathbf{u}|\mathbf{X}]=\mathbf{0}$，在此情况下，首先不需要工具。因此，获得 $\mathrm{E}[\hat{\boldsymbol{\beta}}_{IV}]$ 并不存在简单方法。在证实一致性时不会出现类似问题。于是，得出 $\hat{\boldsymbol{\beta}}_{IV}=\boldsymbol{\beta}+(N^{-1}\mathbf{Z}'\mathbf{X})^{-1}N^{-1}\mathbf{Z}'\mathbf{u}$，其中，$N^{-1}\mathbf{Z}'\mathbf{u}$

项能够脱离 $\mathbf{X}$ 而单独考虑，同时假设 $N^{-1}\mathbf{Z}'\mathbf{u}=\mathbf{0}$ 会产生 $\text{plim}\, N^{-1}\mathbf{Z}'\mathbf{u}=\mathbf{0}$。

因此，我们需要使用其他可供选择的方法来获得 IV 估计量的均值。这里，我们仅仅概述一些重要结果。

起初的研究要做出变量联合正态分布与同方差这种强假设。然后，IV 估计量具有威沙特(Wishart)分布(第 13 章将给出其定义)。令人惊讶的是，甚至在恰好识别情况下，IV 估计量的均值并不存在，这作为有限样本问题存在的信号。如果至少存在一个过度识别约束，那么一定存在均值；如果至少存在两个过度识别，那么一定存在方差。甚至当存在均值时，就相对于 OLS 预测方面的偏倚而言，IV 估计量是有偏的。当拥有更多的过度识别约束时，其偏倚将会增大，最终等于 OLS 估计量的偏倚。戴维森和麦金农(Davidson and Mackinnon，1993，第 221～224 页)曾给出详细讨论。基于幂级数展开的近似也经常得到应用。

是什么决定了有限样本偏倚的大小呢？对于拥有单个回归元 x——内生的并通过简化形式的模型 $x=\mathbf{z}\boldsymbol{\pi}+v$ 而与工具 $\mathbf{z}$ 相关——的回归来说，把集中参数(**concentration parameter**)τ^2 定义成 $\tau^2=\boldsymbol{\pi}'\mathbf{Z}\mathbf{Z}'\boldsymbol{\pi}/\sigma_v^2$。可以证明，IV 的偏倚是 τ^2 的增函数。数量 τ^2/K 是对 $\boldsymbol{\pi}=\mathbf{0}$ 是否成立进行检验的 F 统计量的总体近似，其中，K 表示工具个数。可以证明，统计量 $F-1$ 是 τ^2/K 的近似无偏估计，其中，F 表示第一阶段简化式模型的实际 F 统计量。这就产生了建立在 4.9.2 节给出的 F 统计量基础上的有限样本偏倚的检验。

施泰格和斯托克(Staiger and Stock，1997)在比较弱的分布假设下，获得一些结果。在特殊情况下，不再需要正态条件。他们使用了弱工具渐近特性的方法，即当$N\to\infty$ 时，就一系列具有 τ^2/K 的保持常值的模型而言，获得 IV 估计量的极限分布。在简化模型中，$1/F$ 提供相对于 OLS 而言的 IV 估计量的有限样本偏倚的近似估计。更一般地，给定 F 时偏倚的范围，会随着内生回归元的个数与工具个数而变化。模拟表明，为了确保 IV 中最大偏倚不超过 OLS 偏倚的 10%，我们需要 $F>10$。这个极限值被广泛引用，但是在 6.5 附近却失效，例如，人们对 IV 偏倚为 OLS 偏倚的 20%感到满意。因此，稍欠严格性的经验法则是 $F>5$。谢伊(Shea，1997)已经证明，很小的偏 R^2 与有限样本偏倚也有关系，却不存在类似于使用偏 R^2 作为有限样本偏倚诊断的经验法则。

对于具有一个以上内生回归元的模型，可对每个内生回归元计算各自的 F 统计量。就联合统计量而言，斯托克、赖特和与吾(Stock，Wright and Yogo，2002)提出，利用类似于第一阶段检验 F 统计量的矩阵最小特征值。斯托克和与吾(Stock and Yogo，2003)阐述了，当人们期望的偏倚度、内生变量的个数以及过度识别约束条件的个数变化时，这些特征值的相关临界值。这些表格包括了作为特殊情况的单个内生回归元，同时假定至少两个过度识别约束，所以不能把它们应用于恰好识别模型上。

不仅 IV 估计量可产生有限样本偏倚，而且 IV 标准误差与检验统计量也可产生有限样本偏倚。斯托克等人(Stock et al.，2002)阐述了类似于沃尔德的检验方法，因此，在名义水平 5%下对 $\beta=\beta_0$ 进行检验，比如说，拥有不超过 15% 的真实水平。斯托克和与吾(Stock and Yogo，2003)也提供了关于这一失真方法的详细表

格，内容包括恰好识别模型。

4.9.5 对弱工具响应

在面对弱工具时，实践者要做些什么呢?

正如已提及的，一种方法是限制所用工具个数。这可通过省略一些工具或者对工具加以组合而实现。

如果有限样本偏倚是人们关注的内容，那么其他一些可供选择的估计量可能具有比 2SLS 更好的小样本性质。6.4.4 节将阐述许多其他可供选择的方法，一些方法是 IV 的变形。

尽管强调了有限样本偏倚，但在应用中由弱工具引起的其他问题同样是很重要的。就充分大样本而言，第一阶段的简化式 F 统计量会很大，以至于有限样本偏倚不是什么问题。同时，偏 R^2 可能非常小，甚至对于模型误差与工具之间微小的相关会产生脆弱性。这一点很难加以检验，且难以克服。

正如 4.9.3 节与 4.9.4 节所详述的，估计量在精度上也具有很大的损失。在这些情况下，或者需要更大的样本，或者一定要用可供选择的方法去估计因果边际效应。这些方法在 2.8 节做了概述，而且在本书的其他一些地方也要加以阐述。

4.9.6 IV 应用

克林(Kling, 2001)详细地分析使用靠近学院作为受教育工具的情况。这里，我们使用同样的一组数据，即使用 1976 年 NLS 的年龄在 24～34 岁的 3 010 个男性(数据组)数据，并生成曾先后被卡德(Card, 1995)和克林(Kling, 2001)使用的表格 1。所估计的模型为：

$$\ln w_i = \alpha + \beta_1 s_i + \beta_2 e_i + \beta_3 e_i^2 + \mathbf{x}_{2i}' \gamma + u_i$$

其中，s 表示受教育年数，e 表示工作经历的年数，e^2 表示经历年数的平方，而 $\mathbf{x}_2$ 表示含有 26 个控制变量的向量，这 26 个控制变量主要是地理指标量和父母受教育的程度。

由于缺少关于能力的数据，所以受教育变量被认为是内生的。另外两个工作经历变量是内生的，因为工作经历被计算成年龄减去受教育年数，再减掉 6，这样做在文献中是很平常的，而且受教育是内生的。所以，至少需要三个工具。

此处，确实使用了三个工具，因此该模型是恰好识别的。第一个工具是 *col4*，表示是否靠近四年制学院的标示变量。这个工具已在 4.8.2 节讨论过。另外两个工具是年龄与年龄平方。这两个工具与经历及经历平方之间存在高度相关，然而，它们可以从工资对数模型中省略掉，因为工作经历会起作用。其余的回归元向量 $\mathbf{x}_2$ 作为自身的工具。

尽管年龄很明显是外生的，但是诸如社会技术这些不可观察量可能和年龄与薪水相关。那么年龄和年龄平方作为工具就有问题了。这一点描述了一般性的观点——不用工具有效性的假设。

表 4.5 给出一些结果。β_1 的 OLS 估计为 0.073，因此，受教育额外增加一年，

会使工资平均提高 7.6%[$=100\times(e^{0.073}-1)$]。一旦省略能力，此估计值关于 β_1 是非一致估计值。由于模型是恰好识别的，所以 IV 估计或者等价的 2SLS 估计是 0.132。当受教育额外增加一年，会引起工资增加 14.1%[$=100\times(e^{0.132}-1)$]。

表 4.5　受教育回报：工具变量估计量[a]

	OLS	IV
受教育(s)	0.073	0.132
	(0.004)	(0.049)
R^2	0.304	0.207
谢伊的偏 R^2	—	0.006
关于 s 的第一阶段 F 统计量	—	8.07

[a] 样本是 3 010 名青年男性。因变量是小时工资对数。给定受教育时的系数与标准误差，没有报告经验、经验平方、26 个控制变量和 1 个截距的估计值。对应于 3 个内生回归元——受教育(s)、经验(e)、经验平方(e^2)，三个工具分别为是否靠近四年制大学的标示变量、年龄、年龄平方。偏 R^2 与第一阶段 F 统计量在检验中均对弱工具诊断给予解释。

IV 估计量不如 OLS 的估计量那样有效。正式检验的确没有拒绝同方差性，但我们仍遵循克林(Kling, 2001)的路线，并使用通常的标准差，该标准差非常接近异方差性稳健的标准误差。$\hat{\beta}_{1,OLS}$的标准差是 0.004，而 $\hat{\beta}_{1,IV}$的标准差是 0.049，后者超过前者的 10 倍。相应的其他两个内生回归元的标准误差相差超过 4 倍，而相应的外生回归元的标准误差相差 1.2 倍左右。R^2 由 0.304 下降到 0.207。

通过 R^2 测量可以证实，这些工具与受教育并不是非常关联的。注意到，一个简单的检验是，通过受教育对所有工具的回归(4.58)得出 $R^2=0.297$，若三个添加的工具被省略，则 $R^2=0.291$，两者相差很小。更正式地，此处的谢伊偏 $R^2=0.006\,4=0.08^2$，由式(4.62)知，可以预测，$\hat{\beta}_{1,IV}$的标准误差将被增大 $12.5=1/0.08$ 倍，非常接近于这里观测到的扩大倍数。这使受教育的 t 统计量由 19.64 减少至 2.68。在许多应用中，这种减少会导致统计量不显著。此外，由 4.9.2 节知，甚至工具 $col4_i$ 与误差项 u_i 之间的微小相关，都将导致 IV 的非一致性。

为了研究有限样本偏倚是否也是一个问题，我们实施受教育对所有工具的回归(4.58)。对三个添加工具的联合显著性加以检验，得出 F 统计量为 8.07，这暗示 IV 的偏倚可能是 OLS 偏倚的 10%或 20%。对于其他两个内生变量的类似回归，得到更大一些的 F 统计量，所以年龄是经历的一个很好的工具。倘若存在三个内生回归元，实际上较好的方法是使用已在 4.9.4 节讨论的斯托克等人(Stock et al., 2002)的方法，虽然这里的问题被限制在受教育上，但是，经历与经历平方的谢伊偏 R^2 分别等于 0.087 6 与 0.013 8，而其第一阶段 F 统计量分别是1 772 与 1 542。

如果可利用添加工具法，那么模型变成过度识别的，此外，可利用标准方法对过度识别约束进行检验(参见 8.4.4 节)。

4.10　应用研究

在所有的标准经济计量学软件包中，对于横截面数据来说，本章的估计方法是

可以实施的，但不是所有的方法都可以完成分位数回归。绝大多数方法都提供稳健标准误差作为选项而不是默认项。

应用方面最困难的估计量是工具变量估计量，因为在许多潜在应用中，很难获得那种与误差项无关而与回归元——或者是被用于工具的回归元——适当相关的工具。这种工具可通过对完全结构模型，譬如联立方程系统加以设定而得到。当今的应用研究强调其他一些可供选择的近似方法，比如自然实验。

4.11 文献注释

本章的结果在许多一年级研究生课本中都曾提及，譬如戴维森和麦金农(Davidson and MacKinnon, 2004)；格林(Greene, 2003)；林文夫(Hayashi, 2000)；约翰斯顿和迪纳尔多(Johnston and diNardo, 1997)；米特尔哈默、贾奇和米勒(Mittelhammer、Judge and Miller, 2000)；鲁德(Ruud, 2000)。本节强调的是具有随机回归元的回归、稳健的标准误差、分位数回归、内生性以及工具变量。

4.2 曼斯基(Manski,1991)在一般情况下，包括了由 4.2 节给出的损失函数形式，并对回归给出了优秀的讨论。

4.3 受教育事例已经得到很好的研究。安格里斯特和克鲁格(Angrist and Kruger, 1999)以及卡德(Card, 1999)都提供了最近综述。

4.4 关于最小二乘的历史，参见斯蒂格勒(Stigler, 1986)。勒让德(Legendre)在 1805 年引进了这一方法。高斯(Gauss)在 1810 年把最小二乘法应用于具有正态分布误差项的线性模型，同时提出计算消元法，在后期工作中，他又提出现在被称为高斯—马尔可夫(Gauss-Markov)定理的命题。在 1887 年，高尔顿(Galton)引入回归的概念，意指在家庭个人特性遗传背景下的均值回复[1](mean-reversion)。关于应用于穷人的和福利可利用性的早期“现代”研究，参见尤尔(Yule, 1897)。建立在线性回归模型的最小二乘估计基础上的统计推断是由费希尔(Fisher)显著发展起来的。归功于怀特(White, 1980a)在艾克(Eicker, 1963)早期工作基础上创立起来的 OLS 估计量的方差矩阵的异方差一致性估计，对微观经济计量学的统计推断产生深远的影响，同时已经被推广到许多场合。

4.6 博斯科维克(Boscovich)在 1757 年提出最小绝对偏差估计量，它早于最小二乘法；参见斯蒂格勒(Stigler, 1986)。肯克和巴西特(Kcenker and Bassett, 1978)引进分位数回归，布基斯基(Buchinsky, 1994)对此给出一个综述。一个更基本的解释由肯克和哈洛克(Koenker and Hallock, 2001)给出。

4.7 在联立方程背景下，为了确保识别，赖特(Wright,1928)最早使用了工具变量估计。另外一个经常引用的早期文献是雷厄瑟尔(Reiersol, 1941)的论文，他使用工具变量方法控制回归元的测量误差。萨根(Sargan, 1958)曾经给出早期 IV 估计经典的处理。斯托克和特勒比(Stock and Trebbi,2003)则提供另外一些早期文献。

[1] 又称为均值复归。——译者注

4.8 工具变量估计在经济计量学教材中得到了阐述，这些教材强调代数推导，而缺少必要的直观性。该方法广泛用于经济计量学，因为得到拥有因果解释的估计值是人们所向往的。

4.9 弱工具问题受到应用研究者譬如纳尔逊和施塔茨(Nelson and Startz, 1990)以及邦德等人(Bound et al., 1995)的关注。在理论研究上，许多开创性工作是由纳加尔(Nagar, 1959)做出的，这也是最著名的工作。这一问题削弱了人们对IV估计的热情，归因于弱工具的小样本偏倚则是当今非常活跃的研究专题。一些结果均假定iid正态误差项，并把分析限制在对单一内生回归元的讨论上。邦德等人(Bound et al., 2002)的综述提供许多强调弱工具渐近特性的文献，并且简要考察对非线性模型的推广。哈恩和豪斯曼(Hahn and Hausman, 2003b)的综述阐述其他一些方法及结果，这些内容我们在这里没有给予评述。最近，关于标准误差偏倚的研究工作，参见邦德和温德迈杰(Bond and Windmeijier, 2002)。对于深思熟虑的应用，参见C. I. 李(C. I. Lee, 2001)。

习 题

4-1 考察线性回归模型 $y_i=\mathbf{x}_i'\boldsymbol{\beta}+u_i$，其中，$\mathbf{x}_i$ 表示非随机回归元，而 u_i 具有零均值，且有如下关系：若 $i=j$，则 $\mathrm{E}[u_iu_j]=\sigma^2$；若 $|i-j|=1$，则 $\mathrm{E}[u_iu_j]=\rho\sigma^2$；若 $|i-j|>1$，则 $\mathrm{E}[u_iu_j]=0$。因而，对于相邻观测值，误差是相关的，否则误差是无关的。以矩阵记号表示，我们有 $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{u}$，其中，$\boldsymbol{\Omega}=\mathrm{E}[\mathbf{uu}']$。就此模型而言，利用4.4节给出的结果解答下述问题。

(a) 证明 $\boldsymbol{\Omega}$ 是一个带状矩阵，并且只有其对角线上元素与第一个非对角线上的元素为非零项；求出这些非零项。

(b) 利用式(4.19)求出 $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ 的渐近分布。

(c) 阐述如何求出不依赖于未知参数的 $\mathrm{V}[\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}]$ 的一致估计。

(d) 通常OLS输出估计值 $s^2(\mathbf{X}'\mathbf{X})^{-1}$ 会是 $\mathrm{V}[\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}]$ 的一致估计吗？

(e) 这里，$\mathrm{V}[\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}]$ 的怀特异方差性稳健估计值是一致的吗？

4-2 假定我们估计模型 $y_i=\mu+u_i$，其中，$u_i\sim\mathcal{N}[0,\sigma_i^2]$。

(a) 证明 μ 的OLS估计量可简化为 $\hat{\mu}=\bar{y}$。

(b) 由此直接求出 $\hat{\mu}$ 的方差一致估计值。证明它等于由式(4.21)给出的方差的怀特异方差一致估计值。

4-3 假定数据生成过程是 $y_i=\beta_0x_i+u_i$，$u_i=x_i\varepsilon_i$，$x_i\sim\mathcal{N}[0,1]$，并且 $\varepsilon_i\sim\mathcal{N}[0,1]$。假定数据对于不同 i 是独立的，同时 x_i 与 ε_i 是独立的。注意，$\mathcal{N}[0,\sigma^2]$ 的前四阶中心矩分别是 0、σ^2、0 以及 $3\sigma^4$。

(a) 证明误差项 u_i 是条件异方差的。

(b) 求 $\mathrm{plim}\,N^{-1}\mathbf{X}'\mathbf{X}$。(提示：求 $\mathrm{E}[x_i^2]$，并应用大数定律。)

(c) 求 $\sigma_0^2=\mathrm{V}[u_i]$，其中，期望是关于模型中所有随机变量。

(d) 求 $\mathrm{plim}\,N^{-1}\mathbf{X}'\boldsymbol{\Omega}_0\mathbf{X}=\lim N^{-1}\mathrm{E}[\mathbf{X}'\boldsymbol{\Omega}_0\mathbf{X}]$，其中，$\boldsymbol{\Omega}_0=\mathrm{Diag}[\mathrm{V}[u_i|x_i]]$。

(e) 一旦忽略潜在异方差性，利用前面部分的解答，给出 $\sqrt{N}(\hat{\beta}_{\mathrm{OLS}}-\beta_0)$ 的极限

分布中方差矩阵的默认 OLS 结果(4.22)。

(f) 若考虑任何异方差性,给出$\sqrt{N}(\hat{\beta}_{\mathrm{OLS}}-\beta_0)$极限分布中的方差。你的最终解答应该是数值的。

(g) (e)部分与(f)部分的结果差异符合你的先验信念吗?

4-4 考察具有纯量回归元 $y_i=\beta \mathbf{x}_i+u_i$ 的线性回归模型,其数据对于不同的 i 是独立的,尽管误差可能是条件异方差的。

(a) 证明$(\hat{\beta}_{\mathrm{OLS}}-\beta)=(N^{-1}\sum_i x_i^2)^{-1}N^{-1}\sum_i x_i u_i$。

(b) 把柯尔莫哥洛夫大数定律(定理 A.8)应用到 x_i^2 与 $x_i u_i$ 的平均上,证明 $\hat{\beta}_{\mathrm{OLS}}\xrightarrow{p}\beta$。阐述对 x_i 与 u_i 的数据生成过程做出的任何额外假设。

(c) 把林德伯格—利维中心极限定理(定理 A.14)应用到 $x_i u_i$ 的均值上,证明 $N^{-1}\sum_i x_i u_i/N^{-2}\sum_i \mathrm{E}[u_i^2 x_i^2]\xrightarrow{p}\mathcal{N}[0,1]$。阐述对 x_i 与 u_i 的数据生成过程做出的任何额外假设。

(d) 利用乘积极限正态分布法则(定理 A.17)证明,(c)部分蕴含着 $N^{-1/2}\sum_i x_i u_i\xrightarrow{p}\mathcal{N}[0,\lim N^{-1}\sum_i \mathrm{E}[u_i^2 x_i^2]]$。阐述对 x_i 与 u_i 的数据生成过程所做的任何假设。

(e) 把利用式(2.14)得出的结果与乘积极限正态法则(定理 A.17)结合起来,求出 β 的极限分布。

4-5 考察线性回归模型 $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{u}$。

(a) 求出使 $Q(\boldsymbol{\beta})=\mathbf{u}'\mathbf{W}\mathbf{u}$ 最小化的公式,其中,$\mathbf{W}$ 是满秩的。[提示:对于 $f(\mathbf{x})=f(g(\mathbf{x}))=f(\mathbf{z})$,其中,$\mathbf{z}=g(\mathbf{x})$,关于列向量 $\mathbf{x}$ 与 $\mathbf{z}$ 的矩阵微分的链式法则是$\partial f(\mathbf{x})/\partial\mathbf{x}=(\partial\mathbf{z}'/\partial\mathbf{x})\times(\partial f(\mathbf{z})/\partial\mathbf{z})$。]

(b) 证明若 $\mathbf{W}=\mathbf{I}$,则可简化成 OLS 估计量。

(c) 证明若 $\mathbf{W}=\boldsymbol{\Omega}^{-1}$,则给出 GLS 估计量。

(d) 证明若 $\mathbf{W}=\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$,则给出 2SLS 估计量。

4-6 考察模型 $y=\mathbf{x}'\boldsymbol{\beta}+u$ 的 IV 估计(参见 4.8 节),利用具有满秩的 $N\times K$ 阶矩阵 $\mathbf{Z}$ 在恰好识别情况下的工具 $\mathbf{z}$。

(a) 为使 IV 估计量关于 $\boldsymbol{\beta}$是一致的,$\mathbf{z}$ 必须满足什么根本假设?请解释。

(b) 证明已知恰好识别,由式(4.53)定义的 2SLS 估计量简化成由式(4.51)给出的 IV 估计量。

(c) 给出由于 OLS 的非一致性而需要 IV 估计的现实例子,同时设定合适的工具。

4-7 [取自纳尔逊和施塔茨(Nelson and Startz, 1990)] 考察三个方程的模型:$y=\beta x+u$, $x=\lambda u+\varepsilon$, $z=\gamma\varepsilon+v$, 其中,u、ε、v 是相互独立的误差,而且都是 iid 正态的,均值都为 0,其方差分别为 σ_u^2、σ_ε^2 和σ_v^2。

(a) 证明 $\mathrm{plim}(\hat{\beta}_{\mathrm{OLS}}-\beta)=\lambda\sigma_u^2/(\lambda^2\sigma_u^2+\sigma_\varepsilon^2)$。

(b) 证明 $\rho_{xz}^2=\gamma\sigma_\varepsilon^2/(\lambda^2\sigma_u^2+\sigma_\varepsilon^2)(\gamma^2\sigma_\varepsilon^2+\sigma_v^2)$。

(c) 证明 $\hat{\beta}_{\mathrm{IV}}=m_{zy}/m_{zx}=\beta+m_{zu}/(\lambda m_{zu}+m_{z\varepsilon})$, 其中,$m_{zy}=\sum_i z_i y_i$。

(d) 证明当 γ (或者 ρ_{xz}) $\to 0$ 时,$\hat{\beta}_{\mathrm{IV}}-\beta\to 1/\lambda$。

(e) 证明当 $m_{zu}\to -\gamma\sigma_\varepsilon^2/\lambda$ 时,$\hat{\beta}_{\mathrm{IV}}-\beta\to\infty$。

(f) 当工具不好时，上述两个结果关于 $\hat{\beta}_{IV}-\beta$ 的有限样本偏倚矩意味着什么呢？

4-8 选取 4.6.4 节关于健康支出对数(y)与总支出对数 (x) 数据的 50%随机子样本。

(a) 求 OLS 估计值，同时把斜率系数的通常标准误差与怀特标准误差加以比较。

(b) 求中位数回归估计值，同时把该估计值与 OLS 估计值加以比较。

(c) 求 $q=0.25$ 与 $q=0.75$ 的分位数回归估计值。

(d) 利用你对(a)部分～(c)部分的结果，重新画图 4.2。

4-9 选取 4.9.6 节关于工资与受教育数据的 50%随机子样本，重新绘制表 4.5，并提供适当的解释。

5 极大似然法与非线性最小二乘法估计

5.1 引论

非线性估计量是一个关于因变量的非线性函数的估计量。除了第4章已经阐述的线性回归模型的OLS与IV估计量之外,微观经济计量学中使用的大部分估计量都是非线性估计量。非线性形式可以由许多方式产生。条件均值关于参数可以是非线性的。即使条件均值关于参数是线性的,损失也可能导致非线性估计量。虽然最初模型具有关于参数为线性的条件均值,但删失与截取同样会产生非线性估计量。

我们在这里阐述非线性估计量的基本统计推断结果。对于非线性估计量来说,可以利用的小样本结果是非常有限的。相反,统计推断却建立在应用于大样本的渐近理论基础之上。微观经济计量学广泛使用的统计量都是一致的且渐近正态的。

研究生引论课程中给出的线性回归模型研究的重要内容与渐近理论有两点矛盾。首先,对于大部分非线性估计量来说,由于不存在直接公式,所以需要一些可供选择的证明方法。其次,渐近分布通常可能在最弱的分布假设下获得。这种违背已在4.4节中介绍过,使得对OLS估计量进行异方差性稳健推断成为可能。在这种较弱的假设下,由简单回归方法报告的默认标准误差都是无效的。然而,有些内容需要小心慎重,因为这些最弱的假设能导致估计量自身的非一致性,而这是一个更加根本性的问题。

这里的阐述尽可能是解释性的。大多数教科书都阐述依概率分布和依分布收敛的定义、大数定律(LLN)以及中心极限定理(CLT),而本书把这些专题内容归入附录A中。应用研究者极少关注对一致性与渐近正态性的正式证明。然而,常见情况是,数据应用与最新的或复杂的统计问题遭遇冲突,以至于需要阅读最近的经济计量期刊文章。于是,熟悉一致性与渐近正态性的证明是非常有益的,尤其是在得到估计量方差矩阵的可能形式之前获得好的想法。

5.2节提供一个重要结果的概览。5.3节给出了关于最大化或最小化任何目标函数的极值估计量的更正式研究。建立在估计方程基础上的估计量将在5.4节中加以定义并阐述。5.5节简要阐述建立在稳健标准误差基础上的统计推断,而

完整研究则参考第 7 章。极大似然估计与准极大似然估计将在 5.6 节和 5.7 节加以阐述。非线性最小二乘估计则在 5.8 节给出。5.9 节提供一个详细的例子。

其他的重要参数估计方法，即广义矩方法和非线性工具变量法，将在第 6 章单独研究。

5.2 非线性估计量概览

本节提供非线性估计量的渐近性质的一个概述，更为严谨的研究则由 5.3 节给出，而且本节将阐述对非线性模型中的回归系数加以解释的方法。对于理解后面几章要阐述的横截面和面板数据模型，这些内容极为基础。

5.2.1 *泊松回归例子*

介绍非线性估计的一种特定例子是有益的。这里，我们考察泊松回归，更详细的分析则在第 20 章。

泊松分布适合于因变量 y 仅仅取值为非负整数值 0，1，2，…的情况。它用于对事件发生次数的建模，诸如厂商申请的专利数以及个体就诊次数。

具有速率参数的泊松密度，或更正式地讲，泊松概率质量函数是：

$$f(y|\lambda)=e^{-\lambda}\lambda^{y}/y!,\quad y=0,1,2,\cdots$$

可以证明，$\mathrm{E}[y]=\lambda$ 且 $\mathrm{V}[y]=\lambda$。

建立一个回归模型，即对参数 λ 进行设定，以使具有回归元 $\mathbf{x}$ 以及参数向量 $\boldsymbol{\beta}$ 的特定函数随个体而变化。通常，泊松模型设定为：

$$\lambda=\exp(\mathbf{x}'\boldsymbol{\beta})$$

它具有确保均值 $\lambda>0$ 的优点。因此，对于单个观测值而言，泊松回归模型(**Poisson regression model**)的密度是：

$$f(y|\mathbf{x},\boldsymbol{\beta})=e^{-\exp(\mathbf{x}'\boldsymbol{\beta})}\exp(\mathbf{x}'\boldsymbol{\beta})^{y}/y! \tag{5.1}$$

考察建立在样本$\{(y_i,\ \mathbf{x}_i),\ i=1,\cdots,N\}$基础上的极大似然估计。极大似然估计量[**maximum likelihood (ML) estimator**]是针对对数似然函数求最大值(参见 5.6 节)。似然函数是一个联合密度，即已知独立观测值是单个密度的乘积 $\Pi_i f(y_i|\mathbf{x}_i,\boldsymbol{\beta})$，其中，我们以回归元 $\mathbf{x}$ 为条件。那么，对数似然函数是乘积的对数，它等于对数之和，或者为 $\sum_i \ln f(y_i|\mathbf{x}_i,\boldsymbol{\beta})$。

对于泊松密度(5.1)，第 i 个观测值的对数密度是：

$$\ln f(y_i|\mathbf{x}_i,\boldsymbol{\beta})=-\exp(\mathbf{x}_i'\boldsymbol{\beta})+y_i\mathbf{x}_i'\boldsymbol{\beta}-\ln y_i!$$

因此，泊松 ML 估计量 $\hat{\boldsymbol{\beta}}$ 极大化：

$$Q_N(\boldsymbol{\beta})=\frac{1}{N}\sum\nolimits_{i=1}^{N}\{-\exp(\mathbf{x}_i'\boldsymbol{\beta})+y_i\mathbf{x}_i'\boldsymbol{\beta}-\ln y_i!\} \tag{5.2}$$

其中，由于标度因子是 $1/N$，当 $N\rightarrow\infty$时，$Q_N(\boldsymbol{\beta})$仍为有限的。泊松 ML 估计量是

一阶条件$\partial Q_N(\boldsymbol{\beta})/\partial\boldsymbol{\beta}|_{\hat{\boldsymbol{\beta}}}=\mathbf{0}$的解，或是下式的解：

$$\frac{1}{N}\sum_{i=1}^{N}(y_i-\exp(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{x}_i\Big|_{\hat{\boldsymbol{\beta}}}=\mathbf{0} \tag{5.3}$$

对于式(5.3)中的$\hat{\boldsymbol{\beta}}$来说，不存在显式解。但是，可用数值方法计算$\hat{\boldsymbol{\beta}}$，这将由第10章给出。不过，本章只关注于得到的估计值$\hat{\boldsymbol{\beta}}$的统计性质。

5.2.2 m 估计量

更一般地讲，我们把$q\times 1$维参数向量$\boldsymbol{\theta}$的m估计量(**m-estimator**)$\hat{\boldsymbol{\theta}}$定义为对$N$个子函数的和或均值的目标函数求解最大值：

$$Q_N(\boldsymbol{\theta})=\frac{1}{N}\sum_{i=1}^{N}q(y_i,\mathbf{x}_i,\boldsymbol{\theta}) \tag{5.4}$$

其中，$q(\cdot)$表示纯量函数，y_i表示因变量，$\mathbf{x}_i$表示回归元向量，并且本节的结果均假定，对于不同的i都是独立的。

为了简单起见，把y_i写成纯量形式，但其结果可被推广到向量$\mathbf{y}_i$上，从而包括多变量、面板数据以及方程组形式。用N标记目标函数的下标，表示目标函数依赖于样本数据。本书自始至终用q表示$\boldsymbol{\theta}$的维数。注意，这里的q还用于表示式(5.4)中的子函数$q(\cdot)$。

一旦利用对应$q(y,\mathbf{x},\boldsymbol{\theta})$的特定函数形式，许多经济计量学中的估计量与模型就都是m估计量。重要的例子是极大似然[参见后面的式(5.39)]以及非线性最小二乘法(NLS)[参见后面的式(5.67)]。对式(5.2)求极大值的泊松ML估计量是式(5.4)满足$\boldsymbol{\theta}=\boldsymbol{\beta}$且$q(y,\mathbf{x},\boldsymbol{\beta})=-\exp(\mathbf{x}'\boldsymbol{\beta})+y\mathbf{x}'\boldsymbol{\beta}-\ln y!$的例子。

我们关注估计量$\hat{\boldsymbol{\theta}}$，可把$\hat{\boldsymbol{\theta}}$作为与之相关的一阶条件$\partial Q_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\hat{\boldsymbol{\theta}}}=\mathbf{0}$的解，或者是等价的下式的解：

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial q(y_i,\mathbf{x}_i,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\Big|_{\hat{\boldsymbol{\theta}}}=\mathbf{0} \tag{5.5}$$

这是关于q个未知数的q个方程组，通常$\hat{\boldsymbol{\theta}}$没有显式解。

m估计量这一术语归功于休伯(Humber，1967)，它是极大似然估计量(**maximum likelihood estimator**)的缩略语。许多经济计量学作者，包括雨宫(Amemiya，1985，第105页)、格林(Greene，2003，第461页)以及伍德里奇(Wooldridge，2002，第344页)，都把m估计值定义为对如同式(5.4)的一些项之和求最优化。而其他一些作者，包括谢尔夫林(Serfling，1980)，则把m估计量定义为譬如方程(5.5)的解。休伯(Huber，1967)曾经考虑到这两种情况，休伯(Huber，1981，第43页)以两种显性方式定义m估计量。在本书中，我们称前者为m估计量，而称后者为估计方程估计量(将在5.4节单独阐述)。

5.2.3 m 估计量的渐近性质

一个估计量的重要而令人满意的渐近性质是，估计量是一致的，而且具有至少在大样本下实施统计推断的渐近分布。

一致性

确定 $\hat{\boldsymbol{\theta}}$ 性质的第一步是,准确定义 $\hat{\boldsymbol{\theta}}$ 所要估计的内容。我们假定存在唯一的 $\boldsymbol{\theta}$ 值,记为 $\boldsymbol{\theta}_0$,称为真实参数值(**true parameter value**),由它生成数据。这个识别条件(参见 2.5 节)既需要对数据生成过程的成分进行正确设定,又需要此种表述方式的唯一性。因而,对于泊松例子来说,假定数据生成过程是具有泊松参数 $\exp(\mathbf{x}'\boldsymbol{\beta}_0)$ 的,同时,$\mathbf{x}$ 满足 $\mathbf{x}'\boldsymbol{\beta}^{(1)}=\mathbf{x}'\boldsymbol{\beta}^{(2)}$ 当且仅当 $\boldsymbol{\beta}^{(1)}=\boldsymbol{\beta}^{(2)}$。

对真实参数值而言,带有下标 0 的正式记号广泛应用于第 5 章~第 8 章。出于不同的目的,可以对 $\boldsymbol{\theta}$ 取许多不同值,但我们关注的是两个特殊值——真值 $\boldsymbol{\theta}_0$ 与估计值 $\hat{\boldsymbol{\theta}}$。

即使在大样本中,估计值 $\hat{\boldsymbol{\theta}}$ 也永远不能准确等于 $\boldsymbol{\theta}_0$,这源于样本的内在随机性。然而,我们需要 $\hat{\boldsymbol{\theta}}$ 关于 $\boldsymbol{\theta}_0$ 是一致的(参见附录 A 中的定义 A.2),这意味着 $\hat{\boldsymbol{\theta}}$ 必须依概率收敛(**converge in probability**)到 $\boldsymbol{\theta}_0$,记为 $\hat{\boldsymbol{\theta}}\xrightarrow{p}\boldsymbol{\theta}_0$。

严格建立 m 估计量的一致性是困难的。正式结果由 5.3.2 节给出,而有用的非正式条件由 5.3.7 节给出。ML 估计量与 NLS 估计量的专门化研究将在后面几节给出。

极限正态分布

已知一致性,当 $N\rightarrow\infty$ 时,估计量 $\hat{\boldsymbol{\theta}}$ 在 $\boldsymbol{\theta}_0$ 处具有全部质量的分布。对于 OLS,我们为了获得当 $N\rightarrow\infty$ 时具有非退化分布的随机变量,通过乘以 $\sqrt{N}$ 来放大或重新标度 $\hat{\boldsymbol{\theta}}$。那么,统计推断被处理成假定 N 对渐近理论来说足够大,以便提供良好的近似,但又不要太大,以使 $\hat{\boldsymbol{\theta}}$ 在 $\boldsymbol{\theta}_0$ 处重叠。

因此,我们考察 $\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$ 的特性。对大部分估计量来说,其具有的有限样本分布太复杂,以致不能用于推断。不过,运用渐近理论,可获得当 $N\rightarrow\infty$ 时这种分布的极限。对于微观经济计量学中的大部分估计量来说,这一极限是多变量正态分布。更正式地讲,$\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$ 依分布收敛(**converge in distribution**)到多变量正态分布,依分布收敛已在附录 A 中定义。

回顾 4.4 节,OLS 估计量能表述为:

$$\sqrt{N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)=\left(\frac{1}{N}\sum\nolimits_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\frac{1}{\sqrt{N}}\sum\nolimits_{i=1}^{N}\mathbf{x}_iu_i$$

而其极限分布可通过右边第一项的概率极限与第二项的极限正态分布来获得。通过类似方法,得到 m 估计量的极限分布。我们在 5.3.3 节证明,作为式(5.5)的解估计量总可以被写成:

$$\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)=-\left(\frac{1}{N}\sum\nolimits_{i=1}^{N}\left.\frac{\partial^2 q_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}^+}\right)^{-1}\frac{1}{\sqrt{N}}\sum\nolimits_{i=1}^{N}\left.\frac{\partial q_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}_0}\qquad(5.6)$$

其中,对于 $\hat{\boldsymbol{\theta}}$ 与 $\boldsymbol{\theta}_0$ 之间的某个 $\boldsymbol{\theta}^+$ 而言,倘若二阶导数及其逆都存在,$q_i(\boldsymbol{\theta})=q(y_i,\mathbf{x}_i,\boldsymbol{\theta})$。这个结果可通过泰勒级数展开式来获得。

在适当假设下,将会得到下述 m 估计量的极限分布(**limit distribution**):

$$\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\xrightarrow{d}\mathcal{N}[\mathbf{0},\ \mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}] \tag{5.7}$$

其中，$\mathbf{A}_0^{-1}$ 表示式(5.6)右边第一项的概率极限，并假定第二项收敛到 $\mathcal{N}[\mathbf{0},\ \mathbf{B}_0]$分布。表 5.1 已经给出 $\mathbf{A}_0$ 与 $\mathbf{B}_0$ 的表达式。

表 5.1　m 估计量的渐近性质

性质[a]	代数公式
目标函数	$Q_N(\boldsymbol{\theta})=N^{-1}\sum_i q(y_i,\ \mathbf{x}_i,\ \boldsymbol{\theta})$ 对 $\boldsymbol{\theta}$ 求最大值
例子	ML：$q_i=\ln f(y_i\mid \mathbf{x}_i,\ \boldsymbol{\theta})$ 表示对数密度
	NLS：$q_i=-(y_i-g(\mathbf{x}_i,\ \boldsymbol{\theta}))^2$ 表示负的误差平方
一阶条件	$\partial Q_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}=N^{-1}\sum_{i=1}^N \partial q(y_i,\ \mathbf{x}_i,\ \boldsymbol{\theta})/\partial\boldsymbol{\theta}\mid_{\hat{\theta}}=\mathbf{0}$
一致性	$\text{plim}\, Q_N(\boldsymbol{\theta})$ 是在 $\boldsymbol{\theta}=\boldsymbol{\theta}_0$ 处最大化吗？
一致性(非正式)	$\mathrm{E}[\partial q(y_i,\ \mathbf{x}_i,\ \boldsymbol{\theta})/\partial\boldsymbol{\theta}\mid_{\theta_0}]=\mathbf{0}$ 对吗？
极限分布	$\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\xrightarrow{d}\mathcal{N}[\mathbf{0},\ \mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}]$
	$\mathbf{A}_0=\text{plim}\, N^{-1}\sum_{i=1}^N \partial^2 q_i(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'\mid_{\theta_0}$
	$\mathbf{B}_0=\text{plim}\, N^{-1}\sum_{i=1}^N \partial q_i/\partial\boldsymbol{\theta}\times\partial q_i/\partial\boldsymbol{\theta}'\mid_{\theta_0}$
渐近分布	$\hat{\boldsymbol{\theta}}\overset{a}{\sim}\mathcal{N}[\boldsymbol{\theta}_0,\ N^{-1}\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}]$
	$\hat{\mathbf{A}}=N^{-1}\sum_{i=1}^N \partial^2 q_i(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'\mid_{\hat{\theta}}$
	$\hat{\mathbf{B}}=N^{-1}\sum_{i=1}^N \partial q_i/\partial\boldsymbol{\theta}\times\partial q_i/\partial\boldsymbol{\theta}'\mid_{\hat{\theta}}$

[a] 极限分布方差与渐近方差估计值都是稳健的三明治形式，该形式假定对于不同的 i 是独立的。有关其他方差估计，参见 5.5.2 节。

渐近正态性

为了从极限分布结果(5.7)中获得 $\hat{\boldsymbol{\theta}}$ 的分布，用$\sqrt{N}$去除(5.7)的左边，用 N 去除方差。那么，有：

$$\hat{\boldsymbol{\theta}}\overset{a}{\sim}\mathcal{N}[\boldsymbol{\theta}_0,\ \mathrm{V}[\hat{\boldsymbol{\theta}}]] \tag{5.8}$$

其中，$\overset{a}{\sim}$意味着“渐近分布”(**asymptotically distributed**)，并用 $\mathrm{V}[\hat{\boldsymbol{\theta}}]$表示 $\hat{\boldsymbol{\theta}}$ 的渐近方差(**asymptotic variance**)，它满足：

$$\mathrm{V}[\hat{\boldsymbol{\theta}}]=N^{-1}\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1} \tag{5.9}$$

对渐近分布术语的完整讨论已由 4.4.4 节给出，而且 A.6.4 节也将给予讨论。

结果(5.9)依赖于未知的真实参数 $\boldsymbol{\theta}_0$。它通过计算下式的估计渐近方差(**estimated asymptotic variance**)而得到：

$$\hat{\mathrm{V}}[\hat{\boldsymbol{\theta}}]=N^{-1}\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1} \tag{5.10}$$

其中，$\hat{\mathbf{A}}$ 与 $\hat{\mathbf{B}}$ 表示 $\mathbf{A}_0$ 与 $\mathbf{B}_0$ 的一致估计值。

然而，许多经济计量学软件的默认输出经常使用较简单的估计 $\hat{\mathrm{V}}[\hat{\boldsymbol{\theta}}]=$

$-N^{-1}\hat{\mathbf{A}}^{-1}$，这仅在某些特殊情况下是有效的。进一步的讨论，包括估计 $\mathbf{A}_0$ 与 $\mathbf{B}_0$ 的各种方法，以及进行假设检验，可参见 5.5 节。

m 估计值的两个重要例子是 ML 与 NLS 估计量。命题 5.5 与 5.6 分别给出这些估计量的正式结果。这些估计量的较简单的渐近分布分别由式(5.48)与式(5.77)给出。

泊松 ML 例子

与其他 ML 估计量一样，如果密度得到正确设定，那么泊松 ML 估计量是一致的。然而，把 5.3.7 节的式(5.25)应用到式(5.3)，可以揭示，一致性的基本条件确实是较弱的条件 $\mathrm{E}[y|\mathbf{x}]=\exp(\mathbf{x}_i'\boldsymbol{\beta}_0)$，即对均值正确设定。对于 5.7 节详述的某些其他情况，ML 估计量对于分布的部分错误设定的类似稳健性是成立的。

对于泊松 ML 估计量来说，$\partial q(\boldsymbol{\beta})/\partial\boldsymbol{\beta}=(y_i-\exp(\mathbf{x}'\boldsymbol{\beta}_0))\mathbf{x}$，从而得到：

$$\mathbf{A}_0=-\operatorname{plim} N^{-1}\sum_i\exp(\mathbf{x}_i'\boldsymbol{\beta}_0)\mathbf{x}_i\mathbf{x}_i'$$

与

$$\mathbf{B}_0=\operatorname{plim} N^{-1}\sum_i\mathrm{V}[y_i\,|\,\mathbf{x}_i]\mathbf{x}_i\mathbf{x}_i'$$

于是，$\hat{\boldsymbol{\beta}}\overset{a}{\sim}\mathcal{N}\left[\boldsymbol{\theta}_0,\ N^{-1}\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}\right]$，其中，$\hat{\mathbf{A}}=-N^{-1}\sum_i\exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}})\mathbf{x}_i\mathbf{x}_i'$，而 $\hat{\mathbf{B}}=N^{-1}\sum_i(y_i-\exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}}))^2\mathbf{x}_i\mathbf{x}_i'$。

如果数据确实服从泊松分布，那么 $\mathrm{V}[y|\mathbf{x}]=\mathrm{E}[y|\mathbf{x}]=\exp(\mathbf{x}'\boldsymbol{\beta}_0)$，由于 $\mathbf{A}_0=-\mathbf{B}_0$，导致可能的简化，因此 $\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}=-\mathbf{A}_0^{-1}$。不过，在大多数的计数数据应用中，有 $\mathrm{V}[y|\mathbf{x}]>\mathrm{E}[y|\mathbf{x}]$，因此，最好不要施加这一约束。

5.2.4 非线性回归的系数解释

估计的主要目标常常是实施预测，而不是去检验回归元的统计显著性。

边际效应

关注内容常常是测算边际效应，即当回归元 $\mathbf{x}$ 变化一个单位时，y 的条件均值变动。

对于线性回归模型来说，$\mathrm{E}[y|\mathbf{x}]=\mathbf{x}'\boldsymbol{\beta}$ 蕴含着 $\partial\mathrm{E}[y|\mathbf{x}]/\partial\mathbf{x}=\boldsymbol{\beta}$，因而可把系数作为边际而直接加以解释。对于非线性回归模型来说，这种解释已不再可行。例如，$\mathrm{E}[y|\mathbf{x}]=\exp(\mathbf{x}'\boldsymbol{\beta})$，那么 $\partial E[y|\mathbf{x}]/\partial\mathbf{x}=\exp(\mathbf{x}'\boldsymbol{\beta})\boldsymbol{\beta}$ 就既是参数的函数，又是回归元的函数，边际效应的大小不仅依赖于 $\boldsymbol{\beta}$，还依赖于 $\mathbf{x}$。

一般回归函数

对于一般回归函数(**general regression function**)：

$$\mathrm{E}[y|\mathbf{x}]=g(\mathbf{x},\ \boldsymbol{\beta})$$

其边际效应随 x 的估计值而变化。

一种习惯做法是，阐述由表 5.2 给出的三个边际效应估计值之一。第一个估计值是对所有个体的边际效应进行平均。第二个估计值是在 $\mathbf{x}=\bar{\mathbf{x}}$ 处计算边际效应。第三个估计值是在设定的特征点 $\mathbf{x}=\mathbf{x}^*$ 处进行计算。例如，$\mathbf{x}^*$ 可表示一个 12 年学龄的女性等。也可以考察一个以上的代表性个体。

表 5.2 边际效应:三种不同的估计值

公式	描述
$N^{-1}\sum_i \partial \mathrm{E}[y_i \mid \mathbf{x}_i]/\partial x_i$	所有个体的平均响应
$\partial \mathrm{E}[y \mid \mathbf{x}]/\partial \mathbf{x}\mid_{\bar{\mathbf{x}}}$	平均个体的响应
$\partial \mathrm{E}[y \mid \mathbf{x}]/\partial \mathbf{x}\mid_{\mathbf{x}^*}$	满足 $\mathbf{x}=\mathbf{x}^*$ 个体的响应

在非线性模型中,这三个测量值是各不相同的。然而,在线性模型中,它们都等于 $\boldsymbol{\beta}$。甚至效应的符号与参数符号是不相关的,对某些 $\mathbf{x}$ 值来说,$\partial \mathrm{E}[y \mid \mathbf{x}]/\partial x_j$ 是正的;而对 $\mathbf{x}$ 的其他值来说,$\partial \mathrm{E}[y \mid \mathbf{x}]/\partial x_j$ 则为负的。在解释非线性模型的系数时,必须相当慎重。

计算机程序及应用研究经常报告这些测量值中的第二个。在边际效应数量的层面上,这样做是有意义的。但是,政策关注通常在于全部效应,即第一个测量值,或者在于代表性个体或群体,即第三个测量值。第一个测量值对函数形式 $g(\cdot)$ 的各种不同的选取会有相对很少的变动,而其他两个测量值能够变动得相当大。人们还可利用直方图或非参数密度估计值来阐述边际效应的全部分布。

单指标模型

考虑单指标模型,譬如设定为:

$$\mathrm{E}[y \mid \mathbf{x}]=g(\mathbf{x}'\boldsymbol{\beta}) \tag{5.11}$$

对回归系数进行直接解释是可行的,因此,通过单指标 $\mathbf{x}'\boldsymbol{\beta}$ 数据与参数,均可进入非线性函数 $g(\cdot)$。那么,非线性的均值是回归元及参数的线性组合的非线性函数。就单指标模型而言,可利用微分法(**calculus methods**)进行计算,第 i 个回归元变化的条件均值的效应为:

$$\frac{\partial \mathrm{E}[y \mid \mathbf{x}]}{\partial x_j}=g'(\mathbf{x}'\boldsymbol{\beta})\beta_j$$

其中,$g'(z)=\partial g(z)/\partial z$。由此可得,由于:

$$\frac{\partial \mathrm{E}[y \mid \mathbf{x}]/\partial x_j}{\partial \mathrm{E}[y \mid \mathbf{x}]/\partial x_k}=\frac{\beta_j}{\beta_k}$$

所以系数比值给出了回归元变化所引起的相对效应,这里,把共同因子 $g'(\mathbf{x}'\boldsymbol{\beta})$ 消掉。因此,如果 β_j 是 β_k 的 2 倍,那么 x_j 变化 1 个单位的效应是 x_k 变化 1 个单位效应的 2 倍。此外,若 $g(\cdot)$ 是单调的(**monotonic**),由此可得,系数的符号(**sighs**)就给出了所有可能 $\mathbf{x}$ 的效应符号。

单指标模型由于它们解释简单而具有一些优点。许多标准的非线性模型,譬如 logit、probit 以及 Tobit 模型,都是单指标形式。此外,对 $g(\cdot)$ 的某些选择,允许给出另外的解释,比如本节稍后考虑的著名指数函数以及 14.3.4 节将分析的逻辑斯蒂(**logistic**)cdf。

有限差分法

我们强调对微分法的使用。然而,有限差分法(**finite difference method**)是通

过比较当 x_j 增加 1 个单位时的条件均值与其未增加时的条件均值来计算边际效应。因而有：

$$\frac{\Delta \mathrm{E}[y|\mathbf{x}]}{\Delta x_j}=g(\mathbf{x}+\mathbf{e}_j,\ \boldsymbol{\beta})-g(\mathbf{x},\ \boldsymbol{\beta})$$

其中，$\mathbf{e}_j$ 表示第 j 个元素值为 1、其他元素值为 0 的向量。

对于线性模型来说，有限差分法与微分法会导致相同的估计效应，因为 $\Delta \mathrm{E}[y|\mathbf{x}]/\Delta x_j=(\mathbf{x}'\boldsymbol{\beta}+\beta_j)-\mathbf{x}'\boldsymbol{\beta}=\beta_j$。然而，对于非线性模型来说，这两种方法却给出不同的边际效应估计值，除非 x_j 的变化是无穷小的。

微分法经常用于分析连续回归元，而有限差分法用于分析整数值回归元，譬如（0，1）指示变量。

指数条件均值

举一个例子，考察对指数条件均值函数的系数解释，因而有 $\mathrm{E}[y|\mathbf{x}]=\exp(\mathbf{x}'\boldsymbol{\beta})$。许多计数模型与持续期限模型都使用指数形式。

经过一些代数运算，得到$\partial \mathrm{E}[y|\mathbf{x}]/\partial x_j=\mathrm{E}[y|\mathbf{x}]\times\beta_j$。因此，参数能够被解释为半弹性的（**semi-elasticities**），即 x_j 变化 1 个单位时条件均值增加了 β_j 倍。例如，如果 $\beta_j=0.2$，那么 x_j 变化 1 个单位会使 $\mathrm{E}[y|\mathbf{x}]$增加 0.2 倍，即增加 20%。

相反，如果使用有限差分法，那么边际效应被计算成 $\Delta \mathrm{E}[y|\mathbf{x}]/\Delta x_j=\exp(\mathbf{x}'\boldsymbol{\beta}+\beta_j)-\exp(\mathbf{x}'\boldsymbol{\beta})=\exp(\mathbf{x}'\boldsymbol{\beta})(e^{\beta_j}-1)$。这不同于微分法结果，除非 β_j 很小，进而 $e^{\beta_j}\simeq 1+\beta_j$。例如，如果 $\beta_j=0.2$，那么增加 22.14%，而不是 20%。

5.3 极值估计量

本节内容可以用作微观经济计量学方面的高级研究生课程。本节阐述极值估计量的一致性与渐近正态性的一些重要结果，极值估计量是指，对目标函数求极小值或极大值这类非常广泛的估计量。其表述形式非常简洁。更完整的认识则需要高等研究，譬如雨宫（Amemiya，1985），这里只是研究基础，或者可在纽韦和麦克法登（Newey and McFadden，1994）中找到。

5.3.1 *极值估计量*

对于单个因变量的横截面分析来说，样本来自 N 个观测值$\{(y_i,\ \mathbf{x}_i),\ i=1,\cdots,N\}$，即因变量 y_i 与回归元的列向量 $\mathbf{x}_i$。若以矩阵记号表示，则样本是$(\mathbf{y},\ \mathbf{X})$，其中，$\mathbf{y}$ 表示 $N\times 1$ 维向量，其第 i 个元素为 y_i，而 $\mathbf{X}$ 表示矩阵，其第 i 行为 $\mathbf{x}_i'$，更完整的定义由 1.6 节给出。

关注焦点是估计 $q\times 1$ 维参数向量 $\boldsymbol{\theta}=[\theta_1,\cdots,\theta_q]'$。$\boldsymbol{\theta}_0$ 值称为真实参数值（**true parameter value**），它是生成数据过程中 $\boldsymbol{\theta}$ 的特殊值，该生成数据的过程称为数据生成过程。

我们考察在 $\boldsymbol{\theta}\in\boldsymbol{\Theta}$ 上对随机目标函数 $Q_N(\boldsymbol{\theta})=Q_N(\mathbf{y},\ \mathbf{X},\ \boldsymbol{\theta})$求极大值的估计量 $\hat{\boldsymbol{\theta}}$，其中，为了记号简单起见，$Q_N(\boldsymbol{\theta})$对数据的依赖性仅仅用下标 N 表示。把这

种估计量称为极值估计量(**extremum estimators**),因为它们求解了极大值和极小值问题。

极值估计量可以是全局极大值(**global maximum**),因而有:

$$\hat{\boldsymbol{\theta}}=\arg\max_{\theta\in\Theta}Q_N(\boldsymbol{\theta}) \tag{5.12}$$

通常,极值估计量是一个局部极大值,被计算成有关一阶条件:

$$\left.\frac{\partial Q_N(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|_{\hat{\theta}}=\mathbf{0} \tag{5.13}$$

的解,其中,$\partial Q_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ 表示 $q\times1$ 维列向量,第 k 个元素为$\partial Q_N(\boldsymbol{\theta})/\partial\theta_k$。强调局部极大值的原因是,它是可以作为渐近正态分布的局部极大值。若 $Q_N(\boldsymbol{\theta})$是全局凹的,则局部极大值与全局极大值是重合的。

极值估计量有两个重要例子。对于本章所考虑的 m 估计量,譬如著名的 ML 与 NLS 估计量,$Q_N(\boldsymbol{\theta})$是样本平均,比如残差平方均。对于广义矩方法估计量来说(参见 6.3 节),$Q_N(\boldsymbol{\theta})$是样本平均值的二次形式。

为了具体起见,讨论关注单方程横截面回归。但是,其结果是相当一般的,并可应用到基于满足本节给出性质的最优化的任何估计量。特别地,对纯量因变量不存在限制,而一些学者用 $\mathbf{z}_i$ 代替(y_i, $\mathbf{x}_i$)。于是,$Q_N(\boldsymbol{\theta})$等于 $Q_N(\mathbf{Z},\boldsymbol{\theta})$而不是 $Q_N(\mathbf{y},\mathbf{X},\boldsymbol{\theta})$。

5.3.2 正式一致性定理

首先,我们考虑 2.5 节引入的参数识别。如果数据的分布或者所关注分布的性质是由 $\boldsymbol{\theta}_0$ 所决定的,而 $\boldsymbol{\theta}$ 的任何其他值会导致不同分布,从直观上讲,参数 $\boldsymbol{\theta}_0$ 是可识别的。例如,在线性回归中,我们需要 $\mathrm{E}[\mathbf{y}|\mathbf{X}]=\mathbf{X}\boldsymbol{\beta}_0$ 同时 $\mathbf{X}\boldsymbol{\beta}^{(1)}=\mathbf{X}\boldsymbol{\beta}^{(2)}$,当且仅当 $\boldsymbol{\beta}^{(1)}=\boldsymbol{\beta}^{(2)}$。

估计方法可以不用识别 $\boldsymbol{\theta}_0$。例如,在估计方法忽略了某些有关回归元的情况下。如果取具有参数 $\boldsymbol{\theta}=\boldsymbol{\theta}_0$ 的数据生成过程目标函数的概率极限在 $\boldsymbol{\theta}=\boldsymbol{\theta}_0$ 处达到唯一极大值,那么我们就说估计方法识别 $\boldsymbol{\theta}_0$。这种识别条件是渐近形式。有限样本中产生的实际估计问题将在第 10 章加以讨论。

一致性以下述方式建立起来。当 $N\rightarrow\infty$时,随机目标函数 $Q_N(\boldsymbol{\theta})$,即 m 估计情况下的平均值,依概率收敛到极限函数,极限函数记为 $Q_0(\boldsymbol{\theta})$,在最简单情况下是非随机的。那么,当 $\boldsymbol{\theta}$ 值互相接近时,$Q_N(\boldsymbol{\theta})$与 $Q_0(\boldsymbol{\theta})$的相对应的(全局的或局部的)极大值就应该出现。因为 $Q_N(\boldsymbol{\theta})$的极大值是由 $\hat{\boldsymbol{\theta}}$ 定义的,由此可见,倘若 $\boldsymbol{\theta}_0$ 使得 $Q_0(\boldsymbol{\theta})$极大化,则 $\hat{\boldsymbol{\theta}}$ 依概率收敛到 $\boldsymbol{\theta}_0$。

显然,一致性与识别是紧密相关的,雨宫(Amemiya, 1985,第 230 页)曾经阐述,一种简单方法是,认为识别意指存在一致估计量。进一步讨论,参见纽韦和麦克法登(Newey and McFadden, 1994, 第 2 124 页),以及戴斯特勒和塞弗(Deistler and Seifer, 1978)。

这种方法的重要应用包括詹里希(Jennrich, 1969)与雨宫(Amemiya, 1973)。雨宫(Amemiya, 1985)以及纽韦和麦克法登(Newey and McFadden, 1994)都曾阐

述相当一般的定理。这些定理需要几个假设，包括光滑性（连续性）以及目标函数导数必须存在，为了确保 $Q_N(\boldsymbol{\theta})$ 收敛到 $Q_0(\boldsymbol{\theta})$ 的数据生成过程假设，还要求 $Q_0(\boldsymbol{\theta})$ 在 $\boldsymbol{\theta}=\boldsymbol{\theta}_0$ 处极大化。各种不同的一致性定理使用了稍微不同的假设。

我们阐述归功于雨宫（Amemiya，1985）的两个一致性定理，一个是关于全局性极大值的，一个是关于局部极大值的。雨宫定理中的记号已被修改，这是因为雨宫（Amemiya，1985）定义的目标函数中没有譬如式（5.4）中的正规化 $1/N$ 因子。

定理 5.1（全局极大值的一致性） ［雨宫（Amemiya，1985，定理 4.1.1）］做出下述假设：

(i) 参数空间 $\boldsymbol{\Theta}$ 是 R^q 的一个紧子集。

(ii) 对于所有的 $\boldsymbol{\theta}\in\boldsymbol{\Theta}$，目标函数 $Q_N(\boldsymbol{\theta})$ 是数据的可测函数，同时 $Q_N(\boldsymbol{\theta})$ 在 $\boldsymbol{\theta}\in\boldsymbol{\Theta}$ 内是连续的。

(iii) $Q_N(\boldsymbol{\theta})$ 依概率一致性收敛到非随机函数 $Q_0(\boldsymbol{\theta})$，并且 $Q_0(\boldsymbol{\theta})$ 在 $\boldsymbol{\theta}_0$ 达到全局唯一极大值。

那么，估计量 $\hat{\boldsymbol{\theta}}=\arg\max_{\theta\in\Theta}Q_N(\boldsymbol{\theta})$ 关于 $\boldsymbol{\theta}_0$ 是一致的，即 $\hat{\boldsymbol{\theta}}\xrightarrow{p}\boldsymbol{\theta}_0$。

条件（iii）中的 $Q_N(\boldsymbol{\theta})$ 依概率一致收敛（**uniform convergence in probability**）到：

$$Q_0(\boldsymbol{\theta})=\operatorname{plim}Q_N(\boldsymbol{\theta}) \tag{5.14}$$

意指 $\sup_{\theta\in\Theta}|Q_N(\boldsymbol{\theta})-Q_0(\boldsymbol{\theta})|\xrightarrow{p}0$。

对于局部极大值来说，一阶导数必须存在，但是后面人们只需考虑 $Q_N(\boldsymbol{\theta})$ 及其在 $\boldsymbol{\theta}_0$ 邻域内的特性。

定理 5.2（局部极大值的一致性） ［雨宫（Amemiya，1985，定理 4.1.2）］做出下述假设：

(i) 参数空间 $\boldsymbol{\Theta}$ 是 R^q 的一个开子集。

(ii) 对于所有的 $\boldsymbol{\theta}\in\boldsymbol{\Theta}$，$Q_N(\boldsymbol{\theta})$ 是数据的可测函数，同时 $\partial Q_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ 存在且在 $\boldsymbol{\theta}_0$ 的某个开邻域内是连续的。

(iii) 在 $\boldsymbol{\theta}_0$ 的某个开邻域内，目标函数 $Q_N(\boldsymbol{\theta})$ 依概率一致收敛到 $Q_0(\boldsymbol{\theta})$，且 $Q_0(\boldsymbol{\theta})$ 在 $\boldsymbol{\theta}_0$ 处达到唯一局部极大值。

那么，$\partial Q_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}=\mathbf{0}$ 的一个解关于 $\boldsymbol{\theta}_0$ 是一致的。

使用定理 5.2 的一个例子，稍后由 5.3.4 节给出。

定理 5.1 中的条件（i）允许局部极大值位于参数空间的边界上，而定理 5.2 中的局部极大值必须位于参数空间的内部。定理 5.2 中的条件（ii）还蕴含着 $Q_N(\boldsymbol{\theta})$ 在 $\boldsymbol{\theta}_0$ 的某个开邻域内的连续性，而 $\boldsymbol{\theta}_0$ 的某个邻域 $N(\boldsymbol{\theta}_0)$ 是开的，当且仅当存在以 $\boldsymbol{\theta}_0$ 为中心的球全部位于 $N(\boldsymbol{\theta}_0)$ 中。在这两个定理中，条件（iii）是根本性条件。不论是 $Q_0(\boldsymbol{\theta})$ 的全局极大值还是局部极大值，都必须在 $\boldsymbol{\theta}=\boldsymbol{\theta}_0$ 处取得。条件（iii）的第二部分提供 $\boldsymbol{\theta}_0$ 具有有意义解释及唯一性的识别条件。

对于局部极大值来说，如果仅仅存在一个局部极大值，那么可直接进行分析。于是，通过 $\partial Q_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\hat{\theta}}=\mathbf{0}$ 唯一地定义了 $\hat{\boldsymbol{\theta}}$。当存在多于一个局部极大值时，该定理直接表明，的确有一个局部极大值是一致的，但无法保证哪一个是一致的。在这种情况下，最好是考虑全局极大值，并应用定理 5.1。参见纽韦和麦克法登

(Newey and McFadden, 1994,第 2 117 页)的讨论。

在反映对目标函数 $Q_N(\boldsymbol{\theta})$的选择即模型设定,与用于获得式(5.14)中 $Q_0(\boldsymbol{\theta})$的(**y**, **X**)真实数据生成过程之间,要做出重要区分。对于某些数据生成过程来说,估计量可能是一致的,而对于其他数据生成过程来说,估计量可能是不一致的。在一些情况下,例如泊松 ML 与 OLS 估计量,倘若条件均值得以正确设定,一致性在广泛数据生成过程下就会产生。而在另一些情况下,一致性则需要较强的数据生成过程假设,譬如对密度的正确设定。

5.3.3 渐近正态性

有关渐近正态性的结果,通常会受限于 $Q_N(\boldsymbol{\theta})$的局部极大值。那么,$\hat{\boldsymbol{\theta}}$ 是式(5.13)的解,一般来讲,解关于 $\hat{\boldsymbol{\theta}}$ 是非线性的,并且没有 $\hat{\boldsymbol{\theta}}$ 的显式解。然而,我们用 $\hat{\boldsymbol{\theta}}$ 的线性函数代替该式左边,只是要使用泰勒级数展开式,然后解出 $\hat{\boldsymbol{\theta}}$。

绝大多数所使用的泰勒定理形式是具有余项的逼近式。这里,我们考察准确的一阶泰勒展开式(**exact first-order Taylor expansion**)。对于可微函数 $f(\cdot)$来说,在 x 与 x_0 之间总存在点 x^+,使得:

$$f(x)=f(x_0)+f'(x^+)(x-x_0)$$

其中,$f'(x)=\partial f(x)/\partial x$ 表示 $f(x)$的导数。这一结果也称为中值定理(**mean value theorem**)。

在当前背景下,具体应用时要进行几种变动。纯量函数 $f(\cdot)$用向量函数 $\mathbf{f}(\cdot)$代替,纯量自变量 x、x_0 和 x^+ 则用向量 $\hat{\boldsymbol{\theta}}$ 、$\boldsymbol{\theta}_0$ 以及 $\boldsymbol{\theta}^+$代替。那么有:

$$\mathbf{f}(\hat{\boldsymbol{\theta}})=\mathbf{f}(\boldsymbol{\theta}_0)+\left.\frac{\partial \mathbf{f}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right|_{\theta^+}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) \tag{5.15}$$

其中,$\partial \mathbf{f}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ 表示矩阵,考察 $\hat{\boldsymbol{\theta}}$ 与 $\boldsymbol{\theta}_0$ 之间的某个未知 $\boldsymbol{\theta}^+$ 的形式,对这一矩阵的每一行来说,$\boldsymbol{\theta}^+$ 都不同[参见纽韦和麦克法登(Newey and McFadden, 1994, 第 2 141 页)]。对于局部极限估计量来说,函数 $\mathbf{f}(\boldsymbol{\theta})=\partial Q_N(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ 已经是一阶导数。那么,利用 $\boldsymbol{\theta}_0$ 附近的准确一阶泰勒级数展开式,可得出:

$$\left.\frac{\partial Q_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\hat{\theta}}=\left.\frac{\partial Q_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\theta_0}+\left.\frac{\partial^2 Q_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right|_{\theta^+}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) \tag{5.16}$$

其中,$\partial^2 Q_N(\boldsymbol{\theta})/\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'$表示 $q\times q$ 阶矩阵,其第(j, k)个元素为$\partial^2 Q_N(\boldsymbol{\theta})/\partial \boldsymbol{\theta}_j\partial \boldsymbol{\theta}_k$,而 $\boldsymbol{\theta}^+$ 表示 $\hat{\boldsymbol{\theta}}$ 与 $\boldsymbol{\theta}_0$ 之间的点。

一阶条件设式(5.16)的左边为 0。设右边为 0,并解出$(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$,得到:

$$\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)=-\left(\left.\frac{\partial^2 Q_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right|_{\theta^+}\right)^{-1}\sqrt{N}\left.\frac{\partial Q_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\theta_0} \tag{5.17}$$

其中,我们用$\sqrt{N}$重新标度,以确保非退化极限分布(下面将进一步讨论)。

结果(5.17)提供了 $\hat{\boldsymbol{\theta}}$ 的一个解。它对 $\hat{\boldsymbol{\theta}}$ 的数值计算并没有什么用处,因为它依赖于 $\boldsymbol{\theta}_0$ 与 $\boldsymbol{\theta}^+$,而这两者都是未知的,但是对于理论分析来说,它却是有用的。特别地,如果可以建立 $\hat{\boldsymbol{\theta}}$ 关于 $\boldsymbol{\theta}_0$ 是一致的,那么未知 $\boldsymbol{\theta}^+$ 依概率收敛到 $\boldsymbol{\theta}_0$,因为它

位于 $\hat{\boldsymbol{\theta}}$ 与 $\boldsymbol{\theta}_0$ 之间且 $\hat{\boldsymbol{\theta}}$ 依概率收敛到 $\boldsymbol{\theta}_0$。

结果(5.17)以一种类似于获得 OLS 估计量极限分布的形式表示 $\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$(参见 5.2.3 节)。我们所需的全部内容就是,假定式(5.17)右边第一项的概率分布及第二项的极限正态分布。

由雨宫(Amemiya, 1985)知道,倘若极值估计量满足局部极大值,就得到下述定理。另外,注意到,雨宫(Amemiya, 1985)曾定义不含有正规化 $1/N$ 的目标函数。而且,雨宫用 limE 而不是 plim 的形式定义 $\mathbf{A}_0$ 与 $\mathbf{B}_0$。

定理 5.3(局部极大值的极限分布) [雨宫(Amemiya, 1985, 定理 4.1.3)]除了前面关于局部极大值一致性定理的假设之外,做出下述假设:

(i) $\partial^2 Q_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$ 存在,并且在 $\boldsymbol{\theta}_0$ 的某个开凸邻域内是连续的。

(ii) 对于任何序列 $\boldsymbol{\theta}^+$, $\partial^2 Q_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'|_{\boldsymbol{\theta}^+}$ 依概率收敛到有限非奇异矩阵:

$$\mathbf{A}_0=\operatorname{plim}\partial^2 Q_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'|_{\boldsymbol{\theta}_0} \tag{5.18}$$

使得 $\boldsymbol{\theta}^+\xrightarrow{p}\boldsymbol{\theta}_0$。

(iii) $\sqrt{N}\partial Q_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\boldsymbol{\theta}_0}\xrightarrow{d}\mathcal{N}[\mathbf{0},\ \mathbf{B}_0]$,其中:

$$\mathbf{B}_0=\operatorname{plim}[N\partial Q_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\times\partial Q_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'|_{\boldsymbol{\theta}_0}] \tag{5.19}$$

那么,极值估计量的概率分布(**limit distribution of the extremum estimator**)是:

$$\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\xrightarrow{d}\mathcal{N}[\mathbf{0},\ \mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}] \tag{5.20}$$

其中,估计量 $\hat{\boldsymbol{\theta}}$ 表示 $\partial Q_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}=\mathbf{0}$ 的一致解。

这个证明可通过直接把极限正态乘积规则(定理 A.17)应用到式(5.17)而得到。注意,证明假定 $\hat{\boldsymbol{\theta}}$ 的一致性已经建立。由表 5.1 给出的 $\mathbf{A}_0$ 与 $\mathbf{B}_0$ 的表达式是针对独立 i 的 $Q_N(\boldsymbol{\theta})=N^{-1}\sum_i q_i(\boldsymbol{\theta})$ 情况的特殊化。

式(5.18)与式(5.19)的概率极限是通过($\mathbf{y}$, $\mathbf{X}$)的数据生成过程而得到的。在一些实际应用中,回归元被假定成非随机的,同时期望只是关于 $\mathbf{y}$ 的。而在另外一些情况下,回归元则被处理成随机的,期望则既可以是关于 $\mathbf{y}$ 的,又可以是关于 $\mathbf{X}$ 的。

5.3.4 *泊松 ML 估计量的渐近性质例子*

在具有随机回归元的外生分层抽样条件下,(y_i, $\mathbf{x}_i$)是 inid 的,没有必要假定 y_i 服从泊松分布,我们正式证明泊松 ML 估计量的一致性与渐近正态分性。

证明一致性(consistency)的重要一步是获得 $Q_0(\boldsymbol{\beta})=\operatorname{plim} Q_N(\boldsymbol{\beta})$,并验证 $Q_0(\boldsymbol{\beta})$ 在 $\boldsymbol{\beta}=\boldsymbol{\beta}_0$ 处达到极大值。对于式(5.1)定义的 $Q_N(\boldsymbol{\beta})$,我们有:

$$\begin{aligned}
Q_0(\boldsymbol{\beta}) &= \operatorname{plim} N^{-1}\sum\nolimits_i\{-e^{\mathbf{x}_i'\boldsymbol{\beta}}+y_i\mathbf{x}_i'\boldsymbol{\beta}-\ln y_i!\}\\
&= \operatorname{plim} N^{-1}\sum\nolimits_i\{-\mathrm{E}[e^{\mathbf{x}_i'\boldsymbol{\beta}}]+\mathrm{E}[y_i\mathbf{x}_i'\boldsymbol{\beta}]-\mathrm{E}[\ln y_i!]\}\\
&= \operatorname{plim} N^{-1}\sum\nolimits_i\{-\mathrm{E}[e^{\mathbf{x}_i'\boldsymbol{\beta}}]+\mathrm{E}[e^{\mathbf{x}_i'\boldsymbol{\beta}_0}\mathbf{x}_i'\boldsymbol{\beta}]-\mathrm{E}[\ln y_i!]\}
\end{aligned}$$

第二个等式假定大数定律可应用于每一项。由于(y_i, $\mathbf{x}_i$)为 inid,所以能应用马尔

可夫大数定律(定理 A. 8),条件是如果第二行给出的每个期望值都存在,并且对于某个$\delta>0$,相应的第$(1+\delta)$阶绝对矩存在,同时定理 A. 8 给出的边条件得到满足。例如,设$\delta=1$,因此,可使用二阶矩。第三行需要数据生成过程使得 $\mathrm{E}[y|\mathbf{x}]=\exp(\mathbf{x}'\boldsymbol{\beta}_0)$的假设。第三行的前两个期望值是关于 $\mathbf{x}$ 的,而 $\mathbf{x}$ 是随机的。注意,$Q_0(\boldsymbol{\beta})$既依赖于 $\boldsymbol{\beta}$,又依赖于 $\boldsymbol{\beta}_0$。一旦对 $\boldsymbol{\beta}$ 求导并假定极限,导数与期望可交换,我们得到:

$$\frac{\partial Q_0(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}=-\lim N^{-1}\sum_i \mathrm{E}[e^{\mathbf{x}_i'\boldsymbol{\beta}}\mathbf{x}_i]+\lim N^{-1}\sum_i \mathrm{E}[e^{\mathbf{x}_i'\boldsymbol{\beta}_0}\mathbf{x}_i]$$

其中,$\mathrm{E}[\ln y!]$关于 $\boldsymbol{\beta}$的导数为 0,因为 $\mathrm{E}[\ln y!]$依赖于 $\boldsymbol{\beta}_0$ 即数据生成过程中的真实参数值,但不依赖于 $\boldsymbol{\beta}$。很明显,在 $\boldsymbol{\beta}=\boldsymbol{\beta}_0$ 处,$\partial Q_0(\boldsymbol{\beta})/\partial\boldsymbol{\beta}=\mathbf{0}$ 且$\partial^2 Q_0(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'=-\lim N^{-1}\mathrm{E}[\exp(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i\mathbf{x}_i']$是负定的,因此,$Q_0(\boldsymbol{\beta})$在 $\boldsymbol{\beta}=\boldsymbol{\beta}_0$ 处达到局部极大值,而由定理 5. 2 知,泊松 ML 估计量是一致的。由于这里 $Q_N(\boldsymbol{\beta})$是全局凹的,所以局部极大值等于全部极大值,并且可利用定理 5. 1 建立起一致性。

就泊松 ML 估计量的渐近正态性而言,对于 $\hat{\boldsymbol{\beta}}$ 与 $\boldsymbol{\beta}_0$ 之间的某个未知 $\boldsymbol{\beta}^+$,利用泊松 ML 估计量一阶条件(5. 3)的准确的一阶泰勒级数展开式,得到:

$$\sqrt{N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)=-\left[-N^{-1}\sum_i e^{\mathbf{x}_i'\boldsymbol{\beta}^+}\mathbf{x}_i\mathbf{x}_i'\right]^{-1}N^{-1/2}\sum_i(y_i-e^{\mathbf{x}_i'\boldsymbol{\beta}_0})\mathbf{x}_i \tag{5.21}$$

一旦对回归元 $\mathbf{x}$ 做出充分假设,就可对第一项应用马尔可夫大数定律,而且由于$\hat{\boldsymbol{\beta}}\xrightarrow{p}\boldsymbol{\beta}_0$,可利用 $\boldsymbol{\beta}^+\xrightarrow{p}\boldsymbol{\beta}_0$,我们有:

$$-N^{-1}\sum_i e^{\mathbf{x}_i'\boldsymbol{\beta}^+}\mathbf{x}_i\mathbf{x}_i'\xrightarrow{p}\mathbf{A}_0=-\lim N^{-1}\sum_i \mathrm{E}[e^{\mathbf{x}_i'\boldsymbol{\beta}_0}\mathbf{x}_i\mathbf{x}_i'] \tag{5.22}$$

由假设知,式(5. 21)中的第二项是个纯量回归元 x。于是,$X=(y-\exp(x\beta_0))x$ 具有均值 $\mathrm{E}[X]=0$,因为 $\mathrm{E}[y|x]=\exp(x\beta_0)$已经被假定为一致的,而方差 $\mathrm{V}[X]=\mathrm{E}[\mathrm{V}[y|x]x^2]$。如果涉及$(y-\exp(x\beta_0))x$ 的第$(2+\delta)$阶绝对矩边条件得到满足,就应用李雅普诺夫中心极限定律(定理 A. 15)。对满足 $y\geqslant 0$ 的这个例子,假定 y 的第三阶矩存在,即 $\delta=1$ 并且 x 是有界的,就足够了。若应用中心极限定律,得到:

$$Z_N=\frac{\sum_i(y_i-e^{\beta_0 x_i})x_i}{\sqrt{\sum_i \mathrm{E}[\mathrm{V}[y_i|x_i]x_i^2]}}\xrightarrow{d}\mathcal{N}[0,\ 1]$$

所以:

$$N^{-1/2}\sum_i(y_i-e^{\beta_0 x_i})x_i\xrightarrow{d}\mathcal{N}\left[0,\ \lim N^{-1}\sum_i \mathrm{E}[\mathrm{V}[y_i|x_i]x_i^2]\right]$$

这里假定渐近方差表达式中的极限存在。利用克拉默—沃尔德方法(Cramer-Wold device),把这一结果推广到向量情况(参见定理 A. 16)。那么:

$$N^{-1/2}\sum_i(y_i-e^{\mathbf{x}_i'\boldsymbol{\beta}_0})\mathbf{x}_i\xrightarrow{d}\mathcal{N}\left[\mathbf{0},\ \mathbf{B}_0=\lim N^{-1}\sum_i \mathrm{E}[\mathrm{V}[y_i|x_i]\mathbf{x}_i\mathbf{x}_i']\right] \tag{5.23}$$

因而,由式(5. 21)得到$\sqrt{N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\xrightarrow{d}\mathcal{N}[\mathbf{0},\ \mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}]$,其中,$\mathbf{A}_0$ 由式(5. 22)定义,而 $\mathbf{B}_0$ 则由式(5. 23)定义。

注意到，对这个特殊的例子，为了使泊松 ML 估计量成为一致的且渐近正态的，$y|\mathbf{x}$ 不必是泊松分布。泊松 ML 估计量一致性的根本假设是，dgp 使得 $\mathrm{E}[y|\mathbf{x}]=\exp(\mathbf{x}'\boldsymbol{\beta}_0)$。

对于渐近正态性，尽管需要另外的较高阶矩存在的假设来允许使用 LLN 与 CLT，但根本假设却是 $\mathrm{V}[y|\mathbf{x}]$存在。如果事实上 $\mathrm{V}[y|\mathbf{x}]=\exp(\mathbf{x}'\boldsymbol{\beta}_0)$，那么 $\mathbf{A}_0=-\mathbf{B}_0$，并且更简单地变成$\sqrt{N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\xrightarrow{d}\mathcal{N}[\mathbf{0},\ -\mathbf{A}_0^{-1}]$。这个 ML 例子的结果可推广到将在 5.7.3 节定义的 LEF 密度类。

5.3.5 一致性与渐近正态性的证明

定理 5.1～定理 5.3 中做出的一些假设是相当一般的，而且不必在每个实际应用中都成立。这些假设需要以类似前面泊松 ML 估计量例子的方式逐条加以验证。这里，我们对 m 估计量拟定一个详细方案。

就一致性而言，重要的一步是得出 $Q_N(\boldsymbol{\theta})$的概率极限。这可通过利用 LLN 来完成，因为对于 m 估计量，$Q_N(\boldsymbol{\theta})$是平均值 $N^{-1}\sum_i q_i(\boldsymbol{\theta})$。dgp 上的各种不同假设会导致对不同 LLN 的应用，而更为根本的是，会得到不同的 $Q_0(\boldsymbol{\theta})$表达式。

渐近正态性除了需要一般性的那些假设之外，还需要 dgp 的假设。特别地，为了获得 $\mathbf{A}_0$，我们要求，dgp 的假设能应用 LLN，并且可以应用 CLT 获得 $\mathbf{B}_0$。

对于 m 估计量，当矩阵$\partial^2 Q_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$的每一个元素都是一个平均值时，因为 $Q_N(\boldsymbol{\theta})$是平均值，LLN 可能验证定理 5.3 的条件(ii)。由 5.3.7 节的非正式一致性条件(5.24)以及有限方差 $\mathrm{E}[N\,\partial Q_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\times\partial Q_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'|_{\theta_0}]$可知，由于$\sqrt{N}\partial Q_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\theta_0}$具有 0 均值，所以 CLT 可能会产生定理 5.3 的条件(iii)。

用于获得估计量的极限分布的特殊 CLT 与 LLN，会随着$(\mathbf{y},\ \mathbf{X})$的 dgp 假设而变化。在所有情况下，因变量是随机的。然而，回归元可能是固定的或随机的，并且在后一种情况下，回归元会表现出时间序列相依性(time-series dependence)。这些问题已在 4.4.7 节对 OLS 加以考察过。

普遍的微观经济计量学假设是，回归元对不同的观测值而言是随机且独立的，这对于全国调查的横截面数据来说是合情合理的。对于简单随机抽样来说，数据$(y_i,\ \mathbf{x}_i)$是 iid 的，进而可使用柯尔莫哥洛夫 LLN 与林德伯格—利维 CLT(定理 A.8 与定理 A.14)。更进一步地，在简单随机抽样(5.18)与(5.19)的条件下，可简化成：

$$\mathbf{A}_0=\mathrm{E}\left[\left.\frac{\partial^2 q(y,\ \mathbf{x},\ \boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\theta_0}\right]$$

与

$$\mathbf{B}_0=\mathrm{E}\left[\left.\frac{\partial q(y,\ \mathbf{x},\ \boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial q(y,\ \mathbf{x},\ \boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right|_{\theta_0}\right]$$

其中，$(y,\ \mathbf{x})$表示单个观测值，期望是关于$(y,\ \mathbf{x})$联合分布的。在许多背景下，都使用这种较简单的记号。

对于分层随机抽样以及固定回归元而言，数据$(y_i,\ \mathbf{x}_i)$是 inid 的，从而需要使

用马尔可夫 LLN 与李雅普诺夫 CLT(定理 A.9 与定理 A.15)。这除了需要在 iid 情况下做出那些假设之外,还需要矩假设。在随机回归元情况下,期望是关于 $(y, \mathbf{x})$ 联合分布的,而在固定回归元情况下,例如在可控实验中,对 $\mathbf{x}$ 水平加以设置,式(5.18)与式(5.19)中的期望是仅仅关于 y 的。

对时间序列而言,假定回归元是随机的,并且假定回归元对不同观测值是相关的,以便得到容纳滞后因变量而必需的这一框架。哈密尔顿(Hamilton, 1994)关注这一情况,怀特(White, 2001a)对此做过大量研究。最简单的处理是把随机变量 $(y, \mathbf{x})$ 限制成平稳分布。然而,如果数据是非平稳的并具有单位根,那么收敛速率可能不再是 $\sqrt{N}$,并且极限分布可能是非正态的。

然而,尽管这些重要概念与理论上的差异是针对(**y**, **x**)随机特性而言的,但对于横截面回归来说,最终的极限定理通常源于定理 5.3 给出的一般形式。

5.3.6 讨论

式(5.20)中的方差矩阵形式被称为三明治形式(**sandwich form**),因为 $\mathbf{A}_0^{-1}$ 与 $\mathbf{A}_0^{-1}$ 之间夹着 $\mathbf{B}_0$。由 4.4.4 节引进的三明治形式,将在 5.5.2 节以更详细的方式加以讨论。

渐近结果能够被推广到非一致估计量上。那么,若 $\boldsymbol{\theta}_0$ 被伪真值 $\boldsymbol{\theta}^*$(**pseudo-true value**)所代替,这里伪真值被定义成使 $Q_0(\boldsymbol{\theta})$ 取局部极大值的那个 $\boldsymbol{\theta}$ 值。这将在 5.7.1 节以更详细的方式对准 ML 估计加以考察。然而,在大多数情况下,估计量是一致的,而在稍后一些章节中,为了简化记号,经常把下标 0 省略。

在前面的结果中,目标函数 $Q_N(\boldsymbol{\theta})$ 最初是通过 $1/N$ 正规化加以定义的,于是 $Q_N(\boldsymbol{\theta})$ 的一阶导数用 $\sqrt{N}$ 正规化,而二阶导数没有被正规化,导致了 $\sqrt{N}$ 一致估计量。在一些情况下,可能需要可供选择的其他正规化,最著名的是具有非平稳趋势的时间序列。

一些结果假定,$Q_N(\boldsymbol{\theta})$ 是连续可微的函数。这就排除了诸如最小绝对偏差等一些估计量,因为 $Q_N(\boldsymbol{\theta})=N^{-1}\sum_i|y_i-\mathbf{x}_i'\beta|$。在这种情况下,一种继续研究的方法是,获得可微逼近函数 $Q_N^*(\boldsymbol{\theta})$,使得 $Q_N^*(\boldsymbol{\theta})-Q_N(\boldsymbol{\theta})\xrightarrow{P}0$,同时把前面的定理应用到 $Q_N^*(\boldsymbol{\theta})$ 上。

获得极限分布的重要步骤是,使用泰勒级数展开式进行线性化。泰勒级数展开式对函数的全局逼近的效果欠佳。由于一致性蕴含着大样本量 $\hat{\boldsymbol{\theta}}$ 接近于 $\boldsymbol{\theta}_0$ 的展开点,所以在统计应用中,当逼近渐近为局部逼近时,它们会发挥很好的作用。利用埃奇沃斯展开式(参见 11.4.3 节),可能得到更精致的渐近理论。自助法(参见第 11 章)是经验研究中实施埃奇沃斯展开式的一种方法。

5.3.7 m 估计量一致性的非正式方法

对实践者来说,与定理 5.1 或定理 5.2 关于一致性的正式证明相比,定理 5.3 的极限正态结果更容易证明。这里,我们阐述一种非正式方法,确定使 m 估计量成为一致的所需分布假设的性质及作用。

对作为局部极大值的 m 估计量来说，一阶条件(5.4)蕴含着对 $\hat{\boldsymbol{\theta}}$ 的选取，以使 $\partial q_i(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\hat{\theta}}$ 的平均值等于 0。从直观上讲，得到关于 $\boldsymbol{\theta}_0$ 的一致估计量的必要条件是，$\partial q(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\theta_0}$ 的平均值极限趋于 0，或者：

$$\operatorname{plim}\frac{\partial Q_N(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\bigg|_{\theta_0}=\lim\frac{1}{N}\sum_{i=1}^{N}\mathrm{E}\left[\frac{\partial q_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\bigg|_{\theta_0}\right]=\mathbf{0} \tag{5.24}$$

其中，第一个等式需要应用大数定理的假设，并且式(5.24)中的期望值是针对($\mathbf{y}$, $\mathbf{X}$)总体 dgp 而取的。倘若任何偏离零的情况都会随 $N\to\infty$而消失，则极限就并不要求准确的等式成立。例如，如果期望等于 $1/N$，那么一致性应该成立。条件(5.24)为实践者提供一种非常有用的检查法。一致性的非正式方法(**informal approach to consistency**)是考察估计量 $\hat{\boldsymbol{\theta}}$ 的一阶条件，同时确定在 $\boldsymbol{\theta}=\boldsymbol{\theta}_0$ 处进行计算时这些极限的期望值是否为零。

甚至不太正式地讲，如果我们考察和式中的分量，那么一致性的根本条件(**essential condition**)是，一般观测值是否有：

$$\mathrm{E}[\partial q(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\theta_0}]=\mathbf{0} \tag{5.25}$$

这个条件为实践者提供非常有用的指南。然而，它既不是必要条件，也不是充分条件。如果式(5.25)中的期望等于 $1/N$，那么还有一种可能，式(5.24)概率极限等于0，因此，条件(5.25)不是必要的。为了认识到它不是充分的，考察利用仅有一个观测值，比如说第一个观测值 y_1，去估计具有均值 μ_0 的 iid 的 y。那么，$\hat{\mu}$是 $y_1-\mu=0$ 的解，并且式(5.25)得以满足。但是，很明显，$y_1\overset{p}{\nrightarrow}\mu_0$，因为单个观测值 y_1 具有不趋于 0 的方差。问题在于，式(5.24)中的 plim 不等于 limE。对一致性的正式证明，需要使用譬如定理 5.1 或定理 5.2 之类的定理。

对于泊松回归，式(5.25)的使用揭示了，一致性的根本条件是对 $y|\mathbf{x}$ 的条件均值的正确设定(参见 5.2.3 节)。类似地，OLS 估计量是 $N^{-1}\sum_i\mathbf{x}_i(y_i-\mathbf{x}_i'\boldsymbol{\beta})=\mathbf{0}$ 的解，因此，由式(5.25)，一致性本质上要求 $\mathrm{E}[\mathbf{x}(y-\mathbf{x}'\boldsymbol{\beta}_0)]=\mathbf{0}$。如同 4.7 节给出的那样，假如 $\mathrm{E}[y|\mathbf{x}]\neq\mathbf{x}'\boldsymbol{\beta}_0$，这个条件就失效，这种情况的发生有许多原因。在另一些例子中，用式(5.25)表示，与要求条件均值正确设定相比，一致性则要求更多的参数假设。

为了把式(5.24)的使用与定理 5.2 的条件(iii)连接起来，注意下述内容：

$\partial Q_0(\boldsymbol{\theta})/\partial\boldsymbol{\theta}=\mathbf{0}$	[定理 5.2 的条件(iii)]
$\Rightarrow\ \partial(\operatorname{plim}Q_N(\boldsymbol{\theta}))/\partial\boldsymbol{\theta}=\mathbf{0}$	[由 $Q_0(\boldsymbol{\theta})$的定义]
$\Rightarrow\ \partial(\lim\mathrm{E}[Q_N(\boldsymbol{\theta})])/\partial\boldsymbol{\theta}=\mathbf{0}$	(因为 LLN $\Rightarrow Q_0=\operatorname{plim}Q_N=\lim\mathrm{E}[Q_N]$)
$\Rightarrow\ \lim\partial E[Q_N(\boldsymbol{\theta})]/\partial\boldsymbol{\theta}=\mathbf{0}$	(极限与微分交换)
$\Rightarrow\ \lim\mathrm{E}[\partial Q_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}]=\mathbf{0}$	(微分与期望交换)

最后的等式是非正式条件(5.24)。然而，为了获得这一结果，需要另外的假设，包括对局部极大值的限制、应用大数定律、极限与微分的可交换性，以及微分与期望(即积分)的可交换性。在纯量情况下，微分与极限进行交换的充分条件是，$\lim_{h\to 0}(\mathrm{E}[Q_N(\theta+h)]-\mathrm{E}[Q_N(\theta)])/h=d\mathrm{E}[Q_N(\theta)]/d\theta$ 均匀地位于θ中。

5.4 估计方程

由 5.3.3 节给出的极限分布的推导，能从局部计值估计量推广到被定义为估计方程解的估计量上，这里的估计方程被设置成平均值为 0。第 6 章将给出几个例子。

5.4.1 *估计方程估计量*

设把 $\hat{\boldsymbol{\theta}}$ 定义为以下 q 个估计方程(**estimating equations**)组的解：

$$\mathbf{h}_N(\hat{\boldsymbol{\theta}}) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{h}(y_i,\ \mathbf{x}_i,\ \hat{\boldsymbol{\theta}}) = \mathbf{0} \tag{5.26}$$

其中，$\mathbf{h}(\cdot)$表示 $q\times 1$ 维向量，并假定对于不同的 i，它是独立的。在稍后的 5.4.2 节，将给出 $\mathbf{h}(\cdot)$的例子。

由于选取 $\hat{\boldsymbol{\theta}}$ 以使 $\mathbf{h}(y,\ \mathbf{x},\ \hat{\boldsymbol{\theta}})$的样本平均值等于 0，所以我们希望 $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$，条件是 $\mathbf{h}(y,\ \mathbf{x},\ \boldsymbol{\theta}_0)$的平均值极限趋于 0，即 $\mathrm{plim}\ \mathbf{h}_N(\boldsymbol{\theta}_0)=\mathbf{0}$。如果要应用 LLN，就要求 $\lim \mathrm{E}[\mathbf{h}_N(\boldsymbol{\theta}_0)]=\mathbf{0}$，或者大致地讲，对于第 i 个观测值，有：

$$\mathrm{E}[\mathbf{h}(y_i,\ \mathbf{x}_i,\ \boldsymbol{\theta}_0)]=\mathbf{0} \tag{5.27}$$

最容易建立一致性的方法是，把式(5.26)推导成 m 估计量的一阶条件。

假定具有一致性，估计方程估计量(**estimating equations estimator**)的极限分布能用与 5.3.3 节关于极值估计量相同的方式来获得。在 $\boldsymbol{\theta}_0$ 点附近取 $\mathbf{h}_N(\boldsymbol{\theta})$的准确一阶泰勒级数展开式，如同式(5.15)具有 $\mathbf{f}(\boldsymbol{\theta})=\mathbf{h}_N(\boldsymbol{\theta})$一样，并令等式右边为 $\mathbf{0}$，然后求解。那么有：

$$\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)=-\left(\left.\frac{\partial \mathbf{h}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right|_{\theta^+}\right)^{-1}\sqrt{N}\,\mathbf{h}_N(\boldsymbol{\theta}_0) \tag{5.28}$$

这就得到下述定理。

定理 5.4(估计方程估计量的极限分布) 假定求解式(5.16)的估计方程估计量关于 $\boldsymbol{\theta}_0$ 是一致的，同时做出下面假设：

(i) $\partial\mathbf{h}_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'$ 存在，且在 $\boldsymbol{\theta}_0$ 的某个开凸邻域内是连续的。

(ii) 对于使得 $\boldsymbol{\theta}^+ \xrightarrow{p} \boldsymbol{\theta}_0$ 的任何序列 $\boldsymbol{\theta}^+$，$\partial\mathbf{h}_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'|_{\theta^+}$ 依概率收敛到有限非奇异矩阵：

$$\mathbf{A}_0 = \mathrm{plim}\left.\frac{\partial \mathbf{h}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right|_{\theta_0} = \mathrm{plim}\ \frac{1}{N}\sum\nolimits_{i=1}^{N}\left.\frac{\partial \mathbf{h}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right|_{\theta_0} \tag{5.29}$$

(iii) $\sqrt{N}\mathbf{h}_N(\boldsymbol{\theta}_0)\xrightarrow{d}\mathcal{N}[\mathbf{0},\ \mathbf{B}_0]$，其中：

$$\mathbf{B}_0 = \mathrm{plim}\ N\mathbf{h}_N(\boldsymbol{\theta}_0)\mathbf{h}_N(\boldsymbol{\theta}_0)' = \mathrm{plim}\ \frac{1}{N}\sum\nolimits_{i=1}^{N}\sum\nolimits_{j=1}^{N}\mathbf{h}_i(\boldsymbol{\theta}_0)\mathbf{h}_j(\boldsymbol{\theta}_0)' \tag{5.30}$$

那么，估计方程估计量的极限分布（**limit distribution of the estimating equations estimator**）是：

$$\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\xrightarrow{d}\mathcal{N}[\mathbf{0},\ \mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0'^{-1}] \tag{5.31}$$

其中，不像极值估计量那样，矩阵 $\mathbf{A}_0$ 可能不是对称的，因为它不再是海赛矩阵。

这个定理可通过对雨宫定理 5.3 加以改进而得到证明。注意，定理 5.4 假定一致性已经建立起来。

戈达姆毕（Godambe，1960）曾经证明，对于以回归元为条件的分析来说，最有效的估计方程估计量是设 $\mathbf{h}_i(\boldsymbol{\theta})=\partial\ln f(y_i|\mathbf{x}_i,\ \boldsymbol{\theta})/\partial\boldsymbol{\theta}$。从而，式（5.26）是 ML 估计量的一阶条件。

5.4.2 *类比原理*

为了激发估计量，类比原理使用了总体条件。曼斯基（Manski，1988a）强调了类比原理作为估计的统一论题的重要性。曼斯基（Manski，1988a，第 6 页）曾提供源自戈德伯格（Goldberger，1968，第 4 页）的下述引文：

> 估计类比原理（**analogy principle**）……认为，总体参数可通过样本统计量来估计，样本统计量在样本中具有的性质与总体中参数具有的性质一样。

类比估计量（**analogue estimators**）是指，通过应用类比原理而获得的估计量。总体矩条件（**population moment conditions**）建议，把估计量作为相应样本矩（**sample moment condition**）的解。

4.2 节已经给出应用类比原理的极值估计量的例子。例如，如果预测目的是对总体中的期望损失求极小值，并可使用误差平方损失，那么回归参数 β 可通过对样本误差平方和求极小值而得以估计。

矩方法估计量也是一个例子。例如，在 iid 情况下，如果总体中 $\mathrm{E}[y_i-\mu]=0$，那么我们通过求解相应的样本矩条件 $N^{-1}\sum_i(y_i-\mu)=0$ 而得到估计量，从而得出样本均值 $\hat{\mu}=\bar{y}$。

估计方程估计量可能被认为是类比估计量。如果式（5.27）在总体中成立，那么 $\boldsymbol{\theta}$ 可通过求解相应的样本矩条件（5.26）加以估计。

在微观经济计量学中，广泛使用估计方程估计量。有关理论被归入广义矩方法（**generalized method of moments**），这将在下一章加以阐述，广义矩方法是针对比参数更广泛的矩条件加以扩展的方法。在应用统计学中，此方法用于广义估计方程（**generalized estimating equations**）。

5.5 统计推断

对假设检验和置信区间的详细研究将由第 7 章给出。这里，我们概括如何利用最普遍方法检验线性约束，包括排除性约束、对估计量可能是非线性的沃尔德检验。若使用渐近理论，则正式结果会导致卡方分布及正态分布，而不是正态性条件下源自线性回归的小样本 F 分布与 t 分布。另外，存在几种一致估计极值估计量

方差矩阵的方法，进而得出标准误差、有关检验的统计量以及可供选择的 p 值。

5.5.1 线性约束的沃尔德假设检验

考察对 h 个线性独立约束譬如 H_0 对 H_a 进行检验，其中：

$$H_0: \mathbf{R}\boldsymbol{\theta}_0 - \mathbf{r} = \mathbf{0}$$
$$H_a: \mathbf{R}\boldsymbol{\theta}_0 - \mathbf{r} \neq \mathbf{0}$$

$\mathbf{R}$ 表示 $h\times q$ 阶常数矩阵，$\mathbf{r}$ 表示 $h\times 1$ 维常数向量。例如，如果 $\boldsymbol{\theta}=[\theta_1, \theta_2, \theta_3]$，那么检验是否存在 $\theta_{10}-\theta_{20}=2$，其中，$\mathbf{R}=[1, -1, 0]$，而 $\mathbf{r}=-2$。

如果 $\mathbf{R}\hat{\boldsymbol{\theta}}-\mathbf{r}$ 即 $\mathbf{R}\boldsymbol{\theta}_0-\mathbf{r}$ 的样本估计值显著地不为 $\mathbf{0}$，那么沃尔德检验就拒绝 H_0。这需要 $\mathbf{R}\hat{\boldsymbol{\theta}}-\mathbf{r}$ 的分布知识。假定 $\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\xrightarrow{d}\mathcal{N}[\mathbf{0}, \mathbf{C}_0]$，其中，由式(5.20)知，$\mathbf{C}_0=\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}$。那么，有：

$$\hat{\boldsymbol{\theta}} \overset{a}{\sim} \mathcal{N}[\boldsymbol{\theta}_0, N^{-1}\mathbf{C}_0]$$

因此，在 H_0 为真的条件下，线性组合满足：

$$\mathbf{R}\hat{\boldsymbol{\theta}}-\mathbf{r} \overset{a}{\sim} \mathcal{N}[\mathbf{0}, \mathbf{R}(N^{-1}\mathbf{C}_0)\mathbf{R}']$$

其中，均值为 0，因为在 H_0 为真的条件下，$\mathbf{R}\boldsymbol{\theta}_0-\mathbf{r}=\mathbf{0}$。

卡方检验

一种方便的做法是通过取二次形式，从多元正态分布变成卡方分布，这就产生了沃尔德统计量(**Wald statistics**)。

$$\mathrm{W}=(\mathbf{R}\hat{\boldsymbol{\theta}}-\mathbf{r})'(\mathbf{R}(N^{-1}\hat{\mathbf{C}})\mathbf{R}')^{-1}(\mathbf{R}\hat{\boldsymbol{\theta}}-\mathbf{r})\xrightarrow{d}\chi^2(h) \tag{5.32}$$

在 H_0 为真的条件下，$\mathbf{R}(N^{-1}\mathbf{C}_0)\mathbf{R}'$ 在线性独立约束的假设下是满秩 h 的，并且 $\hat{\mathbf{C}}$ 是 $\mathbf{C}_0$ 的一致估计量。大的 W 值会导致拒绝，并且在 α 水平时，若 $\mathrm{W}>\chi^2_\alpha(h)$，则拒绝 H_0，否则就不拒绝。

然而，实践者时常使用 F 统计量 $F=\mathrm{W}/h$。于是，推断建立在 $F(h, N-q)$ 分布的基础之上，据此希望，这会提供较好的有限样本逼近。注意，当 $N\to\infty$ 时，h 乘以 $F(h, N)$ 分布收敛于 $\chi^2(h)$。

在获得式(5.32)时，用 $\hat{\mathbf{C}}$ 代替 $\mathbf{C}_0$ 在渐近形式上并没有什么差异，但在有限样本中，不同的 $\hat{\mathbf{C}}$ 将导致 W 的各种不同值。在经典线性回归情况下，这一步骤对应于用 s^2 代替 σ^2。如果误差服从正态分布，那么 W/h 确实服从 F 分布(参见 7.2.1 节)。

单系数检验

关注焦点经常是对单个系数不同于 0 进行检验，比如说第 j 个系数。于是，$\mathbf{R}\boldsymbol{\theta}-\mathbf{r}=\theta_j$ 且 $\mathrm{W}=\hat{\theta}_j^2/(N^{-1}\hat{c}_{jj})$，其中，$\hat{c}_{jj}$ 表示 $\hat{\mathbf{C}}$ 中的第 j 个对角元素。在 H_0 下，当对 W 取平方根时，得到：

$$t=\frac{\hat{\theta}_j}{\mathrm{se}[\hat{\theta}_j]}\xrightarrow{d}\mathcal{N}[0, 1] \tag{5.33}$$

其中，$\mathrm{se}(\hat{\theta}_j)=\sqrt{N^{-1}\hat{c}_{jj}}$表示$\hat{\theta}_j$的渐近标准误差。大的$t$值会导致拒绝，而且与W不同的是，统计量$t$能用于单侧检验。

正式地讲，$\sqrt{W}$是渐近z统计量，但是，我们用记号t表示它，就得出通常的"t统计量"，即估计值被其标准误差去除。在有限样本下，一些统计软件包使用标准正态分布，而另一些统计软件包使用t分布来计算临界值、p值以及置信区间。在有限样本下，这两者都不是完全正确的，除了在误差被假定成正态分布的线性回归这一极为特殊的情况外，t分布是准确的。在无限大样本下，这两者作为t分布会产生相同结果，那么对正态分布而言，则会失败。

5.5.2 *方差矩阵估计*

由于一致估计$\mathbf{A}_0$与$\mathbf{B}_0$的方法有许多，所以存在估计$\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0'^{-1}$的一些可行方法。因此，各种不同的经济计量程序应该给出一样的系数估计，然而，有充足的理由认为，在小样本下可以给出不同的标准误差、t统计量以及p值。决定用哪种方法，取决于实践者以及有关dgp分布假设的威力。

方差矩阵的三明治估计

$\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$的极值分布具有方差矩阵$\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0'^{-1}$。由此可得，$\hat{\boldsymbol{\theta}}$具有渐近方差矩阵$N^{-1}\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0'^{-1}$，这里，因为我们考虑的是$\hat{\boldsymbol{\theta}}$而不是$\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$，故除以$N$。

$\hat{\boldsymbol{\theta}}$的渐近方差的三明治估计值(**sandwich estimate**)是具有形式

$$\hat{\mathrm{V}}[\hat{\boldsymbol{\theta}}]=N^{-1}\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}'^{-1} \tag{5.34}$$

的任何估计值，其中，$\hat{\mathbf{A}}$关于$\mathbf{A}_0$是一致的，而$\hat{\mathbf{B}}$关于$\mathbf{B}_0$是一致的。由于$\hat{\mathbf{B}}$夹在$\hat{\mathbf{A}}^{-1}$与$\hat{\mathbf{A}}'^{-1}$之间，所以称为三明治形式。对于许多估计量来说，$\mathbf{A}$表示海赛矩阵，因此$\mathbf{A}^{-1}$是对称的，但情况未必总是如此。

稳健三明治(**robust sandwich**)估计值是其中一种三明治估计值，即估计值$\hat{\mathbf{B}}$在相对弱假设下关于$\mathbf{B}_0$是一致的。这就导致所谓的稳健三明治误差(**robust standard errors**)。一个重要例子是，OLS估计量的方差矩阵的怀特异方差性一致估计值(参见4.4.5节)。在各种特定背景下，稍后几节将详述，以休伯(Huber，1967)命名的稳健三明治估计值，称为休伯估计值；以艾克与怀特(Eicker-White，1980a，b，1982)命名的，称为艾克—怀特估计值；而在平稳时间序列应用中，以纽韦和韦斯特(Newey-West，1987b)命名的，称为纽韦—韦斯特估计值。

关于A与B的估计

这里，我们阐述$\mathbf{A}_0$与$\mathbf{B}_0$的各种不同估计量，既涉及求解$\mathbf{h}_N(\hat{\boldsymbol{\theta}})=\mathbf{0}$的估计方程估计量，又涉及求解$\partial Q_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\hat{\theta}}=\mathbf{0}$的局部极限估计量。

式(5.29)与式(5.18)中$\mathbf{A}_0$的两种标准估计值，都是海赛(**Hessian**)矩阵估计值：

$$\hat{\mathbf{A}}_{\mathrm{H}}=\left.\frac{\partial\mathbf{h}_N(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right|_{\hat{\theta}}=\left.\frac{\partial^2 Q_N(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\hat{\theta}} \tag{5.35}$$

其中，第二个等式解释了运用海赛术语的由来，而期望海赛(**expected Hessian**)矩阵

估计值是：

$$\hat{\mathbf{A}}_{EH}=\mathrm{E}\left[\frac{\partial\mathbf{h}_N(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right]\bigg|_{\hat{\theta}}=\mathrm{E}\left[\frac{\partial^2 Q_N(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]\bigg|_{\hat{\theta}} \tag{5.36}$$

第一项在解析形式比较简单，并潜在地依赖于较少的分布假设，后者更可能是负定的且可逆的。

对于式(5.30)与式(5.19)中的 $\mathbf{B}_0$ 来说，不可能使用明显的估计值 $N\mathbf{h}_N(\hat{\boldsymbol{\theta}})\times\mathbf{h}_N(\hat{\boldsymbol{\theta}})'$，因为当把 $\hat{\boldsymbol{\theta}}$ 定义成满足 $\mathbf{h}_N(\hat{\boldsymbol{\theta}})=\mathbf{0}$ 时，这等于 0。一种估计是做出潜在的强分布假设，以使：

$$\hat{\mathbf{B}}_E=\mathrm{E}[N\mathbf{h}_N(\boldsymbol{\theta})\mathbf{h}_N(\boldsymbol{\theta})']|_{\hat{\theta}}=\mathrm{E}\left[N\frac{\partial Q_N(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial Q_N(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right]\bigg|_{\hat{\theta}} \tag{5.37}$$

对于具有对不同 i 数据独立的 m 估计量与估计方程估计量来说，弱假设是可能的。于是，式(5.30)简化成：

$$\mathbf{B}_0=\mathrm{E}\left[\frac{1}{N}\sum\nolimits_{i=1}^{N}\mathbf{h}_i(\boldsymbol{\theta})\mathbf{h}_i(\boldsymbol{\theta})'\right]$$

由于独立性蕴含着对于 $i\neq j$，$\mathrm{E}[\mathbf{h}_i\mathbf{h}_j']=\mathrm{E}[\mathbf{h}_i]\mathrm{E}[\mathbf{h}_j']$，给定 $\mathrm{E}[\mathbf{h}_i(\boldsymbol{\theta})]=\mathbf{0}$，这也等于 0。这就产生了外积(**outer product**，记为 **OP**)估计值或 **BHHH** 估计值[以伯恩特、霍尔、霍尔和豪斯曼(Berndt，Hall，Hall and Hausman，1974)命名]：

$$\hat{\mathbf{B}}_{OP}=\frac{1}{N}\sum\nolimits_{i=1}^{N}\mathbf{h}_i(\hat{\boldsymbol{\theta}})\mathbf{h}_i(\hat{\boldsymbol{\theta}})'=\frac{1}{N}\sum\nolimits_{i=1}^{N}\frac{\partial q_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\bigg|_{\hat{\theta}}\frac{\partial q_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\bigg|_{\hat{\theta}} \tag{5.38}$$

$\hat{\mathbf{B}}_{OP}$所需的假设比 $\hat{\mathbf{B}}_E$ 所需的要少一些。

在实际应用中，估计 $\mathbf{B}_0$ 时经常调整自由度(**degrees of freedom adjustment**)，对 $\hat{\mathbf{B}}_{OP}$而言，用$(N-q)$而不是 N 除以式(5.38)，而且类似地，用 $N/(N-q)$乘以式(5.37)中的 $\hat{\mathbf{B}}_E$。在非线性模型中，这种调整会产生较好的有限样本绩效，并且它与对具有同方差误差的 OLS 所做出的自由度调整是相符的。对 $\hat{\mathbf{A}}_H$ 或 $\hat{\mathbf{A}}_{EH}$来说，没有类似的调整。

在满足 $\mathbf{A}_0=-\mathbf{B}_0$ 的特殊情况下，可进行简化。重要的例子是，具有同方差误差的 OLS 或 NLS(参见 5.8.3 节)，以及具有正确设定分布的极大似然法(参见 5.6.4 节)。于是，使用$-\hat{\mathbf{A}}^{-1}$或 $\hat{\mathbf{B}}^{-1}$来估计$\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$的方差。与那些利用三明治形式的方法相比，这些估计值对 dgp 错误设定稍欠稳定性。然而，对 dgp 错误设定可能会另外导致 $\hat{\boldsymbol{\theta}}$ 的非一致性，在此情况下，甚至建立在稳健三明治估计值基础上的推断也将是无效的。

对于 5.2 节的泊松例子来说，有 $\hat{\mathbf{A}}_H=\hat{\mathbf{A}}_{EH}=-N^{-1}\sum_i\exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}})\mathbf{x}_i\mathbf{x}_i'$与 $\hat{\mathbf{B}}_{OP}=(N-q)^{-1}\sum_i(y_i-\exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}}))^2\mathbf{x}_i\mathbf{x}_i'$。如果 $\mathrm{V}[y|\mathbf{x}]=\exp(\mathbf{x}'\boldsymbol{\beta}_0)$，即 $y|\mathbf{x}$ 确实服从泊松分布的情况，那么 $\hat{\mathbf{B}}_E=-[N/(N-q)]\hat{\mathbf{A}}_{EH}$，并且可进行简化。

5.6 极大似然法

ML 估计量在一些估计量中占据着特殊地位。在一致渐近正态的估计量中，

它是最有效的估计量。从教学上讲，它也是重要的，因为非线性回归的诸多方法，比如 m 估计，被看成对最先获得的 ML 估计结果的推广与改变。

5.6.1 *似然函数*

归功于费希尔(Fisher，1922)的似然原理(**likelihood principle**)，是选取对观测到的实际样本的似然求极大值的 $\boldsymbol{\theta}$ 值，并作为参数向量 $\boldsymbol{\theta}_0$ 的估计量。在离散情况下，这一似然是通过概率质量函数所得到的概率；而在连续情况下，似然就是密度。考察离散情况。如果 $\boldsymbol{\theta}$ 的一个值蕴含着观测到的数据发生概率是 0.001 2，而 $\boldsymbol{\theta}$ 的第二个值给出较大的概率 0.001 4，那么 $\boldsymbol{\theta}$ 的第二个值是较好的估计量。

此处，联合概率质量函数或密度 $f(\mathbf{y},\mathbf{X}|\boldsymbol{\theta})$，可以被认为是给定数据($\mathbf{y},\mathbf{X}$)时 $\boldsymbol{\theta}$ 的函数。称这种函数为似然函数(**likelihood function**)，并记为 $\mathrm{L}_N(\boldsymbol{\theta}|\mathbf{y},\mathbf{X})$。对 $\mathrm{L}_N(\boldsymbol{\theta})$ 求极大值等价于对对数似然函数(**log-likelihood function**)

$$\mathcal{L}_N(\boldsymbol{\theta})=\ln \mathrm{L}_N(\boldsymbol{\theta})$$

求极大值。我们取自然对数，是因为在应用中，这会产生具有 N 项之和的目标函数，而不是 N 项之积的目标函数。

条件似然

似然函数 $\mathrm{L}_N(\boldsymbol{\theta})=f(\mathbf{y},X|\boldsymbol{\theta})=f(\mathbf{y},\mathbf{X}|\boldsymbol{\theta})f(\mathbf{X}|\boldsymbol{\theta})$既需要对给定 $\mathbf{X}$ 时 $\mathbf{y}$ 的条件密度加以设定，又需要对 $\mathbf{X}$ 的边际密度进行设定。

然而，估计通常建立在条件似然函数(**conditional likelihood function**)$\mathrm{L}_N(\boldsymbol{\theta})=f(\mathbf{y},\mathbf{X}|\boldsymbol{\theta})$的基础上，因为回归目标是对给定 $\mathbf{X}$ 时 $\mathbf{y}$ 的特性进行建模。如果 $f(\mathbf{y},\mathbf{X}|\boldsymbol{\theta})$与 $f(\mathbf{X})$依赖于参数的互不相交集合，那么这就不是一个约束。当情况如此时，普遍做法是省略从属条件。对于极少数的例外，譬如内生抽样(参见第 3 章和第 24 章)，一致估计要求建立在完全联合密度 $f(\mathbf{y},\mathbf{X}|\boldsymbol{\theta})$而不是条件密度 $f(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})$的基础上。

对于横截面数据来说，观测值($y_i,\mathbf{x}_i$)对不同 i 是独立的，其条件密度函数为 $f(y_i|\mathbf{x}_i,\boldsymbol{\theta})$。那么，由独立性，联合条件密度 $f(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})=\prod_{i=1}^{N}f(y_i|\mathbf{x}_i,\boldsymbol{\theta})$得出(条件)对数似然函数：

$$Q_N(\boldsymbol{\theta})=N^{-1}\mathcal{L}_N(\boldsymbol{\theta})=\frac{1}{N}\sum_{i=1}^{N}\ln f(y_i|\mathbf{x}_i,\boldsymbol{\theta}) \qquad (5.39)$$

这里，我们用 N 去除，因此目标函数是平均值。

通过用向量 $\mathbf{y}_i$ 代替纯量 y_i，同时令 $f(y_i|\mathbf{x}_i,\boldsymbol{\theta})$表示以 $\mathbf{x}_i$ 为条件的 $\mathbf{y}_i$ 的联合密度，就能把结果扩展到多变量数据、方程组以及面板数据上。也可参见 5.7.5 节。

例子

就数据类型的广泛性而言，下述方法用于生成完全参数横截面回归模型。在基础统计学课程中，在因变量 y 为 iid 情况下，首先选择某个分布的一个参数或两个参数(或者在一些极少数情况下，为三个参数)情况加以研究。然后，根据回归元与参数 $\boldsymbol{\theta}$，对一个或两个基本参数加以参数化。

一些广泛使用的分布及参数化已由表 5.3 给出。另外一些分布则由附录 B 给

出，而且阐述了抽取伪随机变量的方法。

表 5.3 极大似然：常用密度

模型	y 的范围	密度 $f(y)$	通用参数化
正态模型	$(-\infty, \infty)$	$[2\pi\sigma^2]^{-1/2}e^{-(y-\mu)^2/2\sigma^2}$	$\mu=\mathbf{x}'\boldsymbol{\beta}$, $\sigma^2=\sigma^2$
贝努利模型	0 或 1	$p^y(1-p)^{1-y}$	logit $p=e^{\mathbf{x}'\boldsymbol{\beta}}/(1+e^{\mathbf{x}'\boldsymbol{\beta}})$
指数模型	$(0, \infty)$	$\lambda e^{-\lambda y}$	$\lambda=e^{\mathbf{x}'\boldsymbol{\beta}}$或 $1/\lambda=e^{\mathbf{x}'\boldsymbol{\beta}}$
泊松模型	$0, 1, 2, \cdots$	$e^{-y}\lambda^y/y!$	$\lambda=e^{\mathbf{x}'\boldsymbol{\beta}}$

对于连续型数据$(-\infty, \infty)$来说，正态分布是一个标准分布。经典线性回归模型设 $\mu=\mathbf{x}'\boldsymbol{\beta}$，且假定 σ^2 是常值。

对于取值为 0 或 1 的离散二值数据来说，其密度总是贝努利的，即一种特殊情况的二项式试验。通常，贝努利概率参数化会产生 logit 模型，这已在表 5.3 中列出。此外，还有 $p=\Phi(\mathbf{x}'\boldsymbol{\beta})$的模型，其中，$\Phi(\cdot)$表示标准正态累积分布函数。这些模型将在第 14 章加以分析。

对于正的连续型数据$(0, \infty)$来说，第 17 章～第 19 章将考察著名的持续期限数据，除了表 5.3 给出的指数模型之外，还经常使用比较丰富的威布尔、伽玛以及对数正态模型。

对于取值为 0，1，2，⋯ 整数值的计数数据来说（参见第 20 章），除 5.2.1 节阐述的泊松模型外，经常使用比较丰富的负二项式。令 $\lambda=\exp(\mathbf{x}'\boldsymbol{\beta})$，这确保了正的条件均值。

对于不完整的可观测数据，使用这些分布的删失或者截取变形。最普遍的例子是删失正态的，称为 Tobit 模型，将在 16.3 节加以阐述。

标准的基于似然的模型，几乎很少通过做出误差项分布的假设来加以设定。相反，它们针对因变量的分布直接设定。在 $y\sim\mathcal{N}[\mathbf{x}'\boldsymbol{\beta}, \sigma^2]$的特殊情况下，我们能等价定义 $y=\mathbf{x}'\boldsymbol{\beta}+u$，其中，误差项 $u\sim\mathcal{N}[0, \sigma^2]$。然而，这依赖于由几个其他分布所共有的正态可加性质。例如，如果 y 服从均值为 $\exp(\mathbf{x}'\boldsymbol{\beta})$的泊松分布，我们总能写成 $y=\exp(\mathbf{x}'\boldsymbol{\beta})+u$，但误差项 u 不再服从人们熟悉的分布。

5.6.2 *极大似然估计量*

极大似然估计量（**maximum likelihood estimator**，记为 MLE）是对（条件）对数似然函数求极大估计量，这显然也是极值估计量。通常，MLE 是求解一阶条件

$$\frac{1}{N}\frac{\partial\mathcal{L}_N(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}=\frac{1}{N}\sum_{i=1}^{N}\frac{\partial\ln f(y_i\mid\mathbf{x}_i,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}=\mathbf{0} \tag{5.40}$$

的局部极大值。更正式地讲，这个估计量是条件 MLE，因为它是建立在给定 $\mathbf{x}$ 时 y 的条件密度基础上，但是，普遍做法是使用比较简单的术语 MLE。

梯度向量$\partial\mathcal{L}_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ 称为得分向量（**score vector**），因为它加总了对数密度的一阶导数，而且当在 $\boldsymbol{\theta}_0$ 处进行计算时，称之为有效得分（**efficient score**）。

5.6.3 信息矩阵等式

倘若密度得以正确设定，并且是针对不依赖于 $\boldsymbol{\theta}$ 的 y 范围值，则可以简化 5.3 节中关于 MLE 的结果。

正则条件

ML 的正则条件是：

$$\mathrm{E}_f\left[\frac{\partial \ln f(y \mid \mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]=\int \frac{\partial \ln f(y \mid \mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f(y \mid \mathbf{x}, \boldsymbol{\theta})=\mathbf{0} \tag{5.41}$$

与

$$-\mathrm{E}_f\left[\frac{\partial^2 \ln f(y \mid \mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]=\mathrm{E}_f\left[\frac{\partial \ln f(y \mid \mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ln f(y \mid \mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right] \tag{5.42}$$

其中，记号 $\mathrm{E}_f[\cdot]$ 明显表示，此期望是针对特定密度 $f(y \mid \mathbf{x}, \boldsymbol{\theta})$ 而取得的。结果 (5.41) 蕴含，得分向量具有期望值 0，而由式 (5.42) 可得出式 (5.41)。

5.6.7 节给出的推导要求 y 不依赖于 $\boldsymbol{\theta}$ 的那些范围值，因此，积分与微分可进行交换。

信息矩阵等式

信息矩阵（**information matrix**）是得分向量外积（**outer product of the score vector**）的期望：

$$\mathcal{I}=\mathrm{E}\left[\frac{\partial \mathcal{L}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \mathcal{L}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right] \tag{5.43}$$

术语信息矩阵用 $\mathcal{I}$ 表示，$\mathcal{I}$ 是 $\partial \mathcal{L}_N(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ 的方差，因为由式 (5.41) 可知，$\partial \mathcal{L}_N(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ 具有零均值。于是，大 $\mathcal{I}$ 的值意味着 $\boldsymbol{\theta}$ 的小变化会导致对数似然的大变化，这就包含了所研究的信息 $\boldsymbol{\theta}$。更准确地讲，数量 $\mathcal{I}$ 称为费希尔信息（**Fisher information**），因为还存在其他可供选择的信息检测式。

如果期望是关于 $f(y \mid \mathbf{x}, \boldsymbol{\theta}_0)$ 的，对对数似然函数 (5.39) 来说，正则条件蕴含着：

$$-\mathrm{E}_f\left[\left.\frac{\partial^2 \mathcal{L}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}_0}\right]=\mathrm{E}_f\left[\left.\frac{\partial \mathcal{L}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \mathcal{L}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}_0}\right] \tag{5.44}$$

关系 (5.44) 称为信息矩阵 (IM) 等式，这蕴含信息矩阵也等于 $-\mathrm{E}[\partial^2 \mathcal{L}_N(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}']$。IM 等式蕴含 $-\mathbf{A}_0=\mathbf{B}_0$，其中，$\mathbf{A}_0$ 与 $\mathbf{B}_0$ 已经在式 (5.18) 与式 (5.19) 中定义过。于是，可对定理 5.3 加以简化，因为 $\mathbf{A}_0^{-1} \mathbf{B}_0 \mathbf{A}_0^{-1}=-\mathbf{A}_0^{-1}=\mathbf{B}_0^{-1}$。

等式 (5.42) 是广义信息矩阵等式的特殊情况（**generalized information matrix equality**）：

$$\mathrm{E}_f\left[\frac{\partial \mathbf{m}(y, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right]=-\mathrm{E}_f\left[\mathbf{m}(y, \boldsymbol{\theta}) \frac{\partial \ln f(y \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right] \tag{5.45}$$

其中，$\mathbf{m}(\cdot)$ 表示具有 $\mathrm{E}_f[\mathbf{m}(y, \boldsymbol{\theta})]=\mathbf{0}$ 的向量矩函数，而期望是关于密度 $f(y \mid \boldsymbol{\theta})$ 的。并且，这一结果已经在 5.6.7 节得到，它被用于第 7 章和第 8 章来获得某些检

验统计量的较简单形式。

5.6.4 ML 估计量的分布

正则条件(5.41)与(5.42)导致5.3节中一般结果的简化。倘若期望是关于$f(y|\mathbf{x},\boldsymbol{\theta}_0)$的，一致性的根本条件是$\mathrm{E}[\partial\ln f(y|\mathbf{x},\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\theta_0}]=\mathbf{0}$。由正则条件(5.41)，这是成立的。因而，如果dgp是$f(y|\mathbf{x},\boldsymbol{\theta}_0)$，也就是说，密度被正确设定，那么MLE关于$\boldsymbol{\theta}_0$是一致的。

对于渐近分布来说，通过IM等式，由于$-\mathbf{A}_0=\mathbf{B}_0$，故可以进行简化，这又一次假定密度被正确设定。

这些结果能汇总成下述命题。

命题5.5(ML估计量的分布) 做出下述假设：

(i) dgp是用作定义似然函数的条件密度$f(y_i|\mathbf{x}_i,\boldsymbol{\theta}_0)$。

(ii) 密度函数$f(\cdot)$满足$f(y,\boldsymbol{\theta}^{(1)})=f(y,\boldsymbol{\theta}^{(2)})$，当且仅当$\boldsymbol{\theta}^{(1)}=\boldsymbol{\theta}^{(2)}$。

(iii) 矩阵：

$$\mathbf{A}_0=\operatorname{plim}\frac{1}{N}\frac{\partial^2\mathcal{L}_N(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\bigg|_{\theta_0} \tag{5.46}$$

存在且是有限非奇异的。

(iv) 对数似然的微分与积分次序能够交换。

那么，ML估计量被定义为一阶条件$\partial N^{-1}\mathcal{L}_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}=\mathbf{0}$之解，它关于$\boldsymbol{\theta}_0$是一致的，而且：

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}-\boldsymbol{\theta}_0)\xrightarrow{d}\mathcal{N}[\mathbf{0},\ -\mathbf{A}_0^{-1}] \tag{5.47}$$

条件(i)表明，条件密度被正确设定；条件(i)与(ii)确保了$\boldsymbol{\theta}_0$是可识别的；条件(iii)类似于OLS估计情况下$\operatorname{plim}N^{-1}\mathbf{X}'\mathbf{X}$上的假设；条件(iv)是正则条件成立所必需的。正如一般情况，概率极限与期望都是关于(**y**, **X**)dgp的，或者是针对**y**的，如果假定回归元是非随机的，或者分析是以**X**为条件的。

5.7节将详细考察条件的放松情形。大多数ML例子满足条件(iv)，但它没有排除诸如区间$[0,\theta]$上一致分布的一些模型，因为在这种情况下，y的范围会随θ而变化。于是，不仅$\mathbf{A}_0\neq-\mathbf{B}_0$，而且全局MLE以不同于$\sqrt{N}$的速率收敛并服从非正态极限分布。例如，参见希拉诺和波特(Hirano and Porter, 2003)。

已知命题5.5，所得到的渐近分布时常被写成：

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}}\overset{a}{\sim}\mathcal{N}\left[\boldsymbol{\theta},-\left(\mathrm{E}\left[\frac{\partial^2\mathcal{L}_N(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]\right)^{-1}\right] \tag{5.48}$$

其中，为了记号简单起见，不需要在$\boldsymbol{\theta}_0$处的计算，我们假定应用LLN，因而定义中的plim算子可用linE代替，然后省略limit。在后面章节将经常使用这一记号。

式(5.48)的右边是克拉默—拉奥下界(Cramer-Rao lower bound，记为CRLB)，由基础统计学课程知道，这个下界是小样本无偏估计量方差的下界。对于大样本来说，这里考察的内容即克拉默—拉奥下界是一致渐近正态估计量的方差矩阵的下界，该估计量在$\boldsymbol{\theta}_0$的紧区间上均匀收敛到$\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$的正态性[参见拉

奥(Rao,1973,第 344～351 页)]。粗略地讲,MLE 因为在$\sqrt{N}$一致估计量中具有最小渐近方差而拥有强的吸引力。这个结果要求对条件密度正确设定的强假设。

5.6.5 威布尔回归例子

举个例子,考察建立在威布尔分布基础上的回归,这经常用于对持续期限数据譬如失业期限的长度进行建模(参见第 17 章)。

威布尔分布的密度是 $f(y)=\gamma\alpha y^{\alpha-1}\exp(-\gamma y^{\alpha})$,其中,$y>0$ 且参数 $\alpha>0$ 而 $\gamma>0$。可以证明,$\mathrm{E}[y]=\gamma^{-1/\alpha}\boldsymbol{\Gamma}(\alpha^{-1}+1)$,其中,$\boldsymbol{\Gamma}(\cdot)$表示伽玛函数。标准的威布尔回归模型是通过设定 $\gamma=\exp(\mathbf{x}'\boldsymbol{\beta})$ 来获得的,在此情况下,$\mathrm{E}[y\mid\mathbf{x}]=\exp(-\mathbf{x}'\boldsymbol{\beta}/\alpha)\boldsymbol{\Gamma}(\alpha^{-1}+1)$。给定在不同 i 上的独立性,对数似然函数是:

$$N^{-1}\mathcal{L}_N(\boldsymbol{\theta})=N^{-1}\sum\nolimits_i\{\mathbf{x}_i'\boldsymbol{\beta}+\ln\alpha+(\alpha-1)\ln y_i-\exp(\mathbf{x}_i'\boldsymbol{\beta})y_i^{\alpha}\}$$

对 $\boldsymbol{\beta}$ 与 α 进行微分,得出一阶条件:

$$N^{-1}\sum\nolimits_i\{1-\exp(\mathbf{x}_i'\boldsymbol{\beta})y_i^{\alpha}\}\mathbf{x}_i=\mathbf{0}$$

$$N^{-1}\sum\nolimits_i\left\{\frac{1}{\alpha}+\ln y_i-\exp(\mathbf{x}_i'\boldsymbol{\beta})y_i^{\alpha}\ln y_i\right\}=0$$

与泊松例子不同,一致性本质上要求对分布正确设定。为了理解这一点,考察 $\boldsymbol{\beta}$ 的一阶条件。非正式条件(5.25)$\mathrm{E}[\{1-\exp(\mathbf{x}'\boldsymbol{\beta})y^{\alpha}\}\mathbf{x}]=\mathbf{0}$ 要求 $\mathrm{E}[y^{\alpha}\mid\mathbf{x}]=\exp(-\mathbf{x}'\boldsymbol{\beta})$,其中,幂数 α 没有被限制成为整数。α 的一阶条件会导致 y 上更为复杂的矩条件出现。

因此,我们继续要求密度实际上是威布尔的,满足 $\gamma=\exp(\mathbf{x}'\boldsymbol{\beta}_0)$ 且 $\alpha=\alpha_0$ 的假设。由于 y 的范围不依赖于参数,所以可应用定理 5.5。那么,由式(5.48)知,威布尔 MLE 服从渐近正态分布,其渐近方差为:

$$\mathrm{V}\begin{bmatrix}\hat{\beta}\\ \hat{\alpha}\end{bmatrix}=\left\{-E\begin{bmatrix}\sum_i-e^{\mathbf{x}_i'\beta_0}y_i^{\alpha_0}\mathbf{x}_i\mathbf{x}_i' & \sum_i-e^{\mathbf{x}_i'\beta_0}y_i^{\alpha_0}\ln(y_i)\mathbf{x}_i\\ \sum_i-e^{\mathbf{x}_i'\beta_0}y_i^{\alpha_0}\ln(y_i)\mathbf{x}_i' & \sum_i d_i\end{bmatrix}\right\}^{-1} \quad (5.49)$$

其中,$d_i=-(1/\alpha_0^2)-e^{\mathbf{x}_i'\beta_0}y_i^{\alpha_0}(\ln y_i)^2$。式(5.49)中矩阵的逆需要通过分块求逆来获得,因为非对角项$\partial^2\mathcal{L}_N(\boldsymbol{\beta},\alpha)/\partial\boldsymbol{\beta}\partial\alpha$ 不具有零期望值。在带有零期望交叉导数 $\mathrm{E}[\partial^2\mathcal{L}_N(\boldsymbol{\beta},\alpha)/\partial\boldsymbol{\beta}\partial\alpha']=\mathbf{0}$ 的模型中可进行简化,诸如具有正态分布误差的回归,在此情况下,信息矩阵称关于 $\boldsymbol{\beta}$与 α 为分块对角的。

5.6.6 MLE 的方差矩阵估计

如同 5.5.2 节曾证明的,存在几种一致估计极值估计量方差矩阵的方法。对于 MLE 来说,如果假定信息矩阵等式成立,那么会产生另外的可能性。于是,$\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}$、$-\mathbf{A}_0^{-1}$ 以及 $\mathbf{B}_0^{-1}$ 都是渐近等价的,它们都是这些数量相应的一致估计值。对 MLE 的详细讨论,已由戴维森和麦金农(Davidson and MacKinnon, 1993, 第 18 章)给出。

三明治估计值是以休伯(Huber, 1967)的名字来命名的,称为休伯估计值;或

以怀特(White, 1982)名字来命名,称为怀特估计值,他们在没有施加信息矩阵等式的条件下研究了 MLE 分布。在理论上,三明治估计值比 $-\hat{\mathbf{A}}^{-1}$ 或 $\hat{\mathbf{B}}^{-1}$ 更为稳健。然而,信息矩阵等式失效的原因则会另外导致 $\hat{\boldsymbol{\theta}}_{ML}$ 的更为基础的非一致性复杂问题,注意到这一点是重要的。这是 5.7 节的主题。

5.6.7 *正则条件的推导*

现在,我们正式推导 5.6.3 节曾经表述的正则条件。为了记号简单起见,均不采用下标 i 与回归元向量。

以推导第一个条件(5.41)开始。对密度进行积分等于 1,即:

$$\int f(y|\boldsymbol{\theta})dy = 1$$

两边对 $\boldsymbol{\theta}$ 进行微分,得到 $\frac{\partial}{\partial\boldsymbol{\theta}}\int f(y|\boldsymbol{\theta})dy = \mathbf{0}$。如果积分范围($y$ 的范围)不依赖于 $\boldsymbol{\theta}$,这蕴含着:

$$\int \frac{\partial f(y|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}dy = \mathbf{0} \tag{5.50}$$

现在,$\partial\ln f(y|\boldsymbol{\theta})/\partial\boldsymbol{\theta} = [\partial f(y|\boldsymbol{\theta})/\partial\boldsymbol{\theta}]/[f(y|\theta)]$ 蕴含着:

$$\frac{\partial f(y|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \frac{\partial\ln f(y|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} f(y|\boldsymbol{\theta}) \tag{5.51}$$

把式(5.51)代入式(5.50),得出:

$$\int \frac{\partial\ln f(y|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} f(y|\boldsymbol{\theta})dy = \mathbf{0} \tag{5.52}$$

倘若期望是关于密度 $f(y|\boldsymbol{\theta})$ 的,这就是式(5.41)。

现在,考察第二个条件(5.42),最初推导更为一般的结果。对于某个(可能向量)函数 $\mathbf{m}(\cdot)$,假定:

$$\mathrm{E}[\mathbf{m}(y,\boldsymbol{\theta})] = \mathbf{0}$$

于是,当期望是关于密度 $f(y|\boldsymbol{\theta})$ 取值时,有:

$$\int \mathbf{m}(y,\boldsymbol{\theta}) f(y|\boldsymbol{\theta})dy = \mathbf{0} \tag{5.53}$$

两边对 $\boldsymbol{\theta}'$ 进行微分,并假定微分与积分是可交换的,则得到:

$$\int\left(\frac{\partial\mathbf{m}(y,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'} f(y|\boldsymbol{\theta}) + \mathbf{m}(y,\boldsymbol{\theta})\frac{\partial f(y|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right)dy = \mathbf{0} \tag{5.54}$$

把式(5.51)代入式(5.54),当期望取值是关于密度 $f(y|\boldsymbol{\theta})$ 的,得出:

$$\int\left(\frac{\partial\mathbf{m}(y,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'} f(y|\boldsymbol{\theta}) + \mathbf{m}(y,\boldsymbol{\theta})\frac{\partial\ln f(y|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'} f(y|\boldsymbol{\theta})\right)dy = \mathbf{0} \tag{5.55}$$

或

$$\mathrm{E}\left[\frac{\partial \mathbf{m}(y,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right]=-\mathrm{E}\left[\mathbf{m}(y,\boldsymbol{\theta})\frac{\partial \ln f(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right] \tag{5.56}$$

正则条件(5.42)是 $\mathbf{m}(y,\boldsymbol{\theta})=\partial \ln f(y|\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ 的特殊情况，从而导致 IM 等式(5.44)。更一般的结果(5.56)则会得出广义的 IM 等式(5.45)。

当积分与微分不能交换时，会出现什么情况呢?起点式(5.50)不再成立，这是因为由微分基本定理知，$\int f(y|\boldsymbol{\theta})dy$ 关于 $\boldsymbol{\theta}$ 的导数在积分范围内包括了反映函数 $\boldsymbol{\theta}$ 存在的其他项。从而，$\mathrm{E}[\partial \ln f(y|\boldsymbol{\theta})/\partial\boldsymbol{\theta}]\neq\mathbf{0}$。

当密度被错误设定时，会出现什么情况呢?于是，式(5.52)仍然成立，但不一定蕴含式(5.41)，这是因为式(5.41)中期望将不再与设定密度 $f(y|\boldsymbol{\theta})$ 有关。

5.7 准极大似然法

把准 MLE $\hat{\boldsymbol{\theta}}_{\mathrm{QML}}$ 定义成如下估计量，对被错误设定的对数似然函数即由于错误设定密度而导致的对数似然函数求极大值。通常，这种错误设定会导致出现非一致估计。

本节阐述准 MLE 的一般性质，随后在某些特殊情况下，对准 MLE 保持一致性展开讨论。

5.7.1 *伪真实值*

原则上讲，任何密度的错误设定都会产生非一致性，进而用 $\mathrm{E}[\partial \ln f(y|\mathbf{x},\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\boldsymbol{\theta}_0}]$ 计算出的期望(参见 5.6.4 节)不再是关于 $f(y|\mathbf{x},\boldsymbol{\theta}_0)$ 的。

通过对 5.3.2 节的一般一致性证明的改进，把准 MLE $\hat{\boldsymbol{\theta}}_{\mathrm{QML}}$ 依概率收敛到伪真实值(**pseudo-true value**)$\boldsymbol{\theta}^*$，$\boldsymbol{\theta}^*$ 定义为：

$$\boldsymbol{\theta}^*=\arg\max_{\theta\in\Theta}(\operatorname{plim} N^{-1}\mathcal{L}_N(\boldsymbol{\theta})) \tag{5.57}$$

概率极限是关于真实 dgp 的。如果真实 dgp 不同于用作构建 $\mathcal{L}_N(\boldsymbol{\theta})$ 所假定的密度 $f(y|\mathbf{x},\boldsymbol{\theta})$，那么通常 $\boldsymbol{\theta}^*\neq\boldsymbol{\theta}_0$ 同准 MLE 是非一致的。

休伯(Huber，1967)以及怀特(White，1982)已经证明，除了以 $\boldsymbol{\theta}^*$ 为中心且 IM 等式不再成立外，准 MLE 的渐近分布类似于 MLE 的情况。那么：

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{\mathrm{QML}}-\boldsymbol{\theta}^*)\xrightarrow{d}\mathcal{N}[\mathbf{0},\ \mathbf{A}^{*-1}\mathbf{B}^*\mathbf{A}^{*-1}] \tag{5.58}$$

其中，$\mathbf{A}^*$ 与 $\mathbf{B}^*$ 均由式(5.18)与式(5.19)定义，只是概率极限是关于未知的真实 dgp 的，并且在 $\boldsymbol{\theta}^*$ 处进行计算。如同 5.5.2 节一样，可获得在 $\hat{\boldsymbol{\theta}}_{\mathrm{QML}}$ 处计算的一致估计值 $\hat{\mathbf{A}}^*$ 与 $\hat{\mathbf{B}}^*$。

如果准 MLE 保持一致性，那么这一分布结果就可用于统计推断。除下一节将给出的解释之外，若准 MLE 是非一致的，则一般来说，$\boldsymbol{\theta}^*$ 没有简单解释。然而，如果关注于估计的准确性，那么式(5.58)还是一个有用的结果。结果(5.58)还提供了怀特信息矩阵检验的动机(参见 8.2.8 节)以及用于参数模型之间进行区别的

翁检验(**Vuong's test**)(参见 8.5.3 节)。

5.7.2 库尔贝克—利布勒距离

回顾 4.2.3 节,若 $\mathrm{E}[y|\mathbf{x}]\neq\mathbf{x}'\boldsymbol{\beta}_0$,则在平方误差损失下,OLS 估计量还能被解释成最佳线性预测式。怀特(White, 1982)曾提出,在性质上类似于准 MLE 的解释。

设 $f(\mathbf{y}|\boldsymbol{\theta})$表示 $y_1,\cdots,y_N$ 的假定联合密度,而设 $h(\mathbf{y})$表示真实密度,它是未知的,为了简单起见,这里没有采用回归元的相依性。把库尔贝克—利布勒信息准则(**Kullback-Leibler information criterion**,记为 **KLIC**)定义成:

$$\mathrm{KLIC}=\mathrm{E}\left[\ln\left(\frac{h(\mathbf{y})}{f(\mathbf{y}|\boldsymbol{\theta})}\right)\right] \tag{5.59}$$

其中,期望是关于 $h(\mathbf{y})$的,当存在 $\boldsymbol{\theta}_0$ 使得 $h(\mathbf{y})=f(\mathbf{y}|\boldsymbol{\theta}_0)$时,KLIC 就取极小值 0,也就是说,密度被正确设定,而且 KLIC 值越大,就越不知道真实密度。

于是,准 MLE $\hat{\boldsymbol{\theta}}_{\mathrm{QML}}$对 $f(\mathbf{y}|\boldsymbol{\theta})$与 $h(\mathbf{y})$之间的距离求极小值,其中,距离是利用 KLIC 测量。为获得这一结果,注意在适当假设下,$\mathrm{plim}\, N^{-1}\mathcal{L}_N(\boldsymbol{\theta})=\mathrm{E}[\ln f(\mathbf{y}|\boldsymbol{\theta})]$。因此,$\hat{\boldsymbol{\theta}}_{\mathrm{QML}}$收敛到使 $\mathrm{E}[\ln f(\mathbf{y}|\boldsymbol{\theta})]$取极大值的 $\boldsymbol{\theta}^*$。然而,由于 $\mathrm{KLIC}=\mathrm{E}[\ln h(\mathbf{y})]-\mathrm{E}[\ln f(\mathbf{y}|\boldsymbol{\theta})]$,同时因为期望是关于 $h(\mathbf{y})$的,第一项不依赖于 $\boldsymbol{\theta}$,这等价于求 KLIC 的极小值。

5.7.3 线性指数族

在一些特殊情况下,甚至当密度被部分错误设定时,准 MLE 仍是一致的。一个众所周知的例子是,倘若 $\mathrm{E}[y|\mathbf{x}]=\mathbf{x}'\boldsymbol{\beta}_0$,尽管误差是非正态的,但具有正态性的线性回归模型的准 MLE 是一致的。泊松 MLE 提供了第二个例子(参见 5.3.4 节)。

类似于错误设定的稳健性,被建立在线性指数族(Linear Exponential family, LEF)中的密度基础上的其他模型所享有。线性指数族密度能够写成:

$$f(y|\mu)=\exp\{a(\mu)+b(\mu)+c(\mu)y\} \tag{5.60}$$

其中,我们已给出 LEF 的均值参数化,因而 $\mu=\mathrm{E}[y]$。可以证明,对这一密度来说,$\mathrm{E}[y]=-[c'(\mu)]^{-1}a'(\mu)$,而 $\mathrm{V}[y]=[c'(\mu)]^{-1}$,其中,$c'(\mu)=\partial c(\mu)/\partial\mu$ 且 $a'(\mu)=\partial a(\mu)/\partial\mu$。各种不同的函数 $a(\cdot)$与 $c(\cdot)$会产生族中的不同密度。可把式(5.60)中的项 $b(y)$正规化为常值,以此保证概率之和或积为 1。密度的剩余部分 $\exp\{a(\mu)+c(\mu)y\}$表示关于 y 为线性的指数函数,因此可解释为线性指数项。

大部分密度不能用这种形式表达。然而,几种重要的密度都是 LEF 密度,包括那些由表 5.4 给出的。由表 5.3 阐述过的这些密度,在表 5.4 中用式(5.60)的形式重新表述。其他的 LEF 密度包括具有已知试验数的二项式(贝努利作为一种特殊情况)、一些负二项式(几何及泊松模型作为一种特殊情况),以及单参数伽玛(指数作为一种特殊情况)。

表 5.4 线性指数族密度:重要例子

分布	$f(y)=\exp\{a(\cdot)+b(y)+c(\cdot)y\}$	E[y]	$\mathrm{V}[y]=[c'(\mu)]^{-1}$
正态分布(σ^2 已知)	$\exp\left\{\frac{-\mu^2}{2\sigma^2}-\frac{1}{2}\ln(2\pi\sigma^2)-\frac{y^2}{2\sigma^2}+\frac{\mu}{\sigma^2}y\right\}$	μ	σ^2
贝努利分布	$\exp\{\ln(1-p)+\ln[p/(1-p)]y\}$	$\mu=p$	$\mu(1-\mu)$
指数分布	$\exp\{\ln\lambda-\lambda y\}$	$\mu=1/\lambda$	μ^2
泊松分布	$\exp\{-\lambda-\ln y!+y\ln\lambda\}$	$\mu=\lambda$	μ

对回归来说,把参数 $\mu=\mathrm{E}[y|\mathbf{x}]$建模成:

$$\mu=g(\mathbf{x},\boldsymbol{\beta}) \tag{5.61}$$

随不同模型而变化的设定函数(参见 5.7.4 节),部分地依赖于对 y 范围的限制,从而依赖于 μ。于是,LEF 对数似然是:

$$\mathcal{L}_N(\boldsymbol{\beta})=\sum_{i=1}^{n}\{a(g(\mathbf{x}_i,\boldsymbol{\beta}))+b(y_i)+c(g(\mathbf{x}_i,\boldsymbol{\beta}))y_i\} \tag{5.62}$$

若利用前面提到的关于 y 的前二阶矩信息,一阶条件可重新表述成为:

$$\frac{\partial\mathcal{L}_N(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}=\sum_{i=1}^{N}\frac{y_i-g(\mathbf{x}_i,\boldsymbol{\beta})}{\sigma_i^2}\times\frac{\partial g(\mathbf{x}_i,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}=\mathbf{0} \tag{5.63}$$

其中,$\sigma_i^2=[c'(g(\mathbf{x}_i,\boldsymbol{\beta}))]^{-1}$被假定成对应于特殊 LEF 密度的方差函数。例如,对贝努利、指数以及泊松而言,σ_i^2 分别等于 $g_i(1-g_i)$、$1/g_i^2$ 以及 g_i,其中,$g_i=g(\mathbf{x}_i,\boldsymbol{\beta})$。

准 MLE 可求解这些方程,但是不再假定 LEF 密度被正确设定。古里耶克斯、蒙福特和特罗农(Gouriéroux, Monfort, and Trognon, 1984a)已经证明,倘若 $\mathrm{E}[y|\mathbf{x}]=g(\mathbf{x},\boldsymbol{\beta}_0)$,则准 MLE $\boldsymbol{\beta}_{\mathrm{QML}}$是一致的。这是一个对一阶条件(5.63)取期望值的清晰形式,如果 $\mathrm{E}[y|\mathbf{x}]=g(\mathbf{x},\boldsymbol{\beta}_0)$,那么它在 $\boldsymbol{\beta}=\boldsymbol{\beta}_0$ 处的计算值作为具有期望值等于 0 的误差 $y-g(\mathbf{x},\boldsymbol{\beta}_0)$ 的加权和。

因此,倘若给定 $\mathbf{x}$ 时,y 的条件均值被正确设定,则基于 LEF 密度的准 MLE 是一致的。注意到,关于 y 的实际 dgp 不必是 LEF。它是一个设定密度,可能被错误设定为 LEF。

然而,甚至对正确条件均值而言,基于方差 $-\mathbf{A}_0^{-1}$,对方差、标准误差和统计量默认输出进行调整是有保证的。一般来讲,应该使用三明治形式 $\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}$,除非给定 $\mathbf{x}$ 时,y 的条件方差也被正确设定,在此情况下,$\mathbf{A}_0=-\mathbf{B}_0$。然而,对于贝努利模型来说,总是有 $\mathbf{A}_0=-\mathbf{B}_0$。可利用式(5.36)与式(5.38)获得一致标准误差。

LEF 是非常特殊的情况。通常,对密度在任何方面的错误设定,都会导致 MLE 的非一致性。甚至在 LEF 情况下,准 MLE 能仅用于预测条件均值,而对正确设定密度而言,它能用于预测条件分布。

5.7.4 广义线性模型

在统计学文献中(参见由麦卡拉和尔德撰写的这方面的文献),把建立在假定 LEF 密度基础上的模型称为广义线性模型。广义线性模型类别是应用统计学中关

于非线性横截面回归的最广泛使用的框架。由表 5.3 知，它包括非线性最小二乘法、泊松模型、几何模型、probit、logit、二项式(已知试验次数)、伽玛以及指数回归模型。我们给出一个简短概述，介绍标准的广义线性模型(GLM)术语。

标准广义线性模型设定(5.61)中的条件均值是较简单的单指标形式，因而 $\mu=g(\mathbf{x}'\boldsymbol{\beta})$。于是，$g^{-1}(\mu)=\mathbf{x}'\boldsymbol{\beta}$，而函数 $g^{-1}(\cdot)$ 称为连接函数(**link function**)。例如，泊松模型的通常设定为对数连接函数，这是因为如果 $\mu=\exp(\mathbf{x}'\boldsymbol{\beta})$，那么 $\ln\mu=\mathbf{x}'\boldsymbol{\beta}$。

一阶条件(5.63)变成 $\sum_i[(y_i-g_i)/c'(g_i)]g_i'\mathbf{x}_i=\mathbf{0}$，其中，$g_i=g(\mathbf{x}_i'\boldsymbol{\beta})$ 且 $g_i'=g'(\mathbf{x}_i'\boldsymbol{\beta})$。选取连接函数以使 $c'(g(\mu))=g'(\mu)$，这样做在计算上有一些优点，进而这些一阶条件简化成 $\sum_i(y_i-g_i)\mathbf{x}_i=\mathbf{0}$，或者误差 (y_i-g_i) 正交于回归元。典型连接函数[1](**canonical link function**)被定义成函数 $g^{-1}(\cdot)$，满足 $c'(g(\mu))=g'(\mu)$ 并且随 $c(\mu)$ 而变化，进而随广义线性模型而变化。典型连接函数对正态而言，使得 $\mu=\mathbf{x}'\boldsymbol{\beta}$；对泊松而言，使得 $\mu=\exp(\mathbf{x}'\boldsymbol{\beta})$；对二值数据而言，使得 $\mu=\exp(\mathbf{x}'\boldsymbol{\beta})/[1+\exp(\mathbf{x}'\boldsymbol{\beta})]$。最后一个式子是由表 5.3 给出的 logit 形式。达到预期的极大似然对数值与拟合对数似然值之差的 2 倍，被称为离差(deviance)，该测量值是将线性回归的平方残差和推广到其他非线性指数族的回归模型上。

建立在 LEF 基础上的模型限制太强，因为所有矩都依赖于唯一一个基本参数 $\mu=g(\mathbf{x}'\boldsymbol{\beta})$。广义线性模型文献通过做出下述方便假设来施加一些另外的结构，即 LEF 方差可能通过纯量倍数 α 被错误设定，因此，$\mathrm{V}[y|\mathbf{x}]=\alpha\times[c'(g(\mathbf{x},\boldsymbol{\beta}))]^{-1}$，其中，必须满足 $\alpha\neq1$。例如，对于泊松分布来说，设 $\mathrm{V}[y|\mathbf{x}]=\alpha g(\mathbf{x},\boldsymbol{\beta})$ 而不是 $g(\mathbf{x},\boldsymbol{\beta})$。已知这种方差错误设定，可以证明 $\mathbf{B}_0=-\alpha\mathbf{A}_0$，因而准 MLE 的方差矩阵是 $-\alpha\mathbf{A}_0^{-1}$，这仅仅需要通过用 α 乘以非三明治 ML 方差矩阵 $-\mathbf{A}_0^{-1}$ 来重新标度。广泛使用的 α 一致估计值是 $\hat{\alpha}=(N-K)^{-1}\sum_i(y_i-\hat{g}_i)^2/\hat{\sigma}_i^2$，其中，$\hat{g}_i=g(\mathbf{x}_i,\hat{\boldsymbol{\beta}}_{\mathrm{QML}})$，$\hat{\sigma}_i^2=c'[(\hat{g}_i)]^{-1}$，用 $(N-K)$ 而不是 N 去除，被认为是对提供小样本更好的估计值。更详细内容，参见前面提及的参考文献以及卡梅伦和特里维迪(Cameron and Trivedi, 1986, 1998)。

许多统计软件都包括广义线性模型模块，倘若 $\mathrm{V}[y|\mathbf{x}]=\alpha[c'(g(\mathbf{x},\boldsymbol{\beta}))]^{-1}$，该模块作为默认而给出正确的标准误差。作为一种可选择的方法，人们能利用 ML 进行估计，其标准误差可用稳健三明治 $\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}$ 公式获得。在实际应用中，三明治误差类似于那些利用简单的广义线性模型修正所获得的误差。然而，另一种估计广义线性模型的方法是通过加权非线性最小二乘法，如同 5.8.6 节末尾所述。

5.7.5 *多因变量的准 MLE*

本章关注纯量因变量，但是该理论还可应用于多元情况。假定因变量 $\mathbf{y}$ 是 $m\times1$ 维向量形式，而数据 $(\mathbf{y}_i, \mathbf{x}_i)$ 对于不同 i 都是独立的，$i=1,\cdots,N$。稍后几章将给出的例子包括看似不相关方程、同一因变量的第 i 个个体样本具有 m 个观测值的面板数据，以及聚类数据，其中，对第 ij 个观测值而言，数据关于 j 的 m 个可

〔1〕 又称标准连接函数。——译者注

能值是相关的。

已知 $f(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})$ 和以 $\mathbf{x}$ 为条件的 $\mathbf{y}=(y_1,\cdots,y_m)$ 的联合密度，那么完全有效 MLE 如同式(5.39)后所注释的，可以求 $N^{-1}\sum_i \ln f(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\theta})$ 极大值。不过，在多元应用中，$\mathbf{y}$ 的联合密度是复杂的。已知 m 个单变量密度 $f_j(y_j|\mathbf{x},\boldsymbol{\theta})$ 的唯一知识，$j=1,\cdots,m$，其中，y_j 表示 $\mathbf{y}$ 的第 j 个分量，得到较简单的估计量是可能的。例如，对于多元计数数据来说，人们从 m 个独立的、每个计数的单变量负二项式密度开始研究，而不是从可能相关的多元计数模型开始。

于是，考察建立在单变量密度之积 $\prod_j f_j(y_j|\mathbf{x},\boldsymbol{\theta})$ 基础上的准 MLE $\hat{\boldsymbol{\theta}}_{\text{QML}}$，它对：

$$Q_N(\boldsymbol{\theta})=\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{m}\ln f(y_{ij}|\mathbf{x}_i,\boldsymbol{\theta}) \tag{5.64}$$

求极大值。伍德里奇(Wooldridge, 2002)把这种估计量称为偏 MLE，这是因为此密度仅仅被部分设定。

偏 MLE 是满足 $q_i=\sum_j \ln f(y_{ij}|\mathbf{x}_i,\boldsymbol{\theta})$ 的 m 估计量。一致性根本条件(5.25)要求，$\mathrm{E}[\sum_j \partial f(y_{ij}|\mathbf{x}_i,\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\theta_0}]=\mathbf{0}$。如果边缘密度 $f(y_{ij}|\mathbf{x}_i,\boldsymbol{\theta}_0)$ 被正确设定，那么这个条件成立，进而由正则条件可知，$\mathrm{E}[\partial f(y_{ij}|\mathbf{x}_i,\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\theta_0}]=\mathbf{0}$。

因而，倘若单变量密度 $f_j(y_j|\mathbf{x},\boldsymbol{\theta})$ 被正确设定，则偏 MLE 是一致的。一致性并不需要 $f(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})=\prod_j f_j(y_j|\mathbf{x},\boldsymbol{\theta})$ 。然而，$y_1,\cdots,y_m$ 的相依性将导致信息矩阵等式失效，因此，标准误差应该利用满足：

$$\mathbf{A}_0=\frac{1}{N}\sum\nolimits_{i=1}^{N}\sum\nolimits_{j=1}^{m}\left.\frac{\partial^2\ln f_{ij}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\theta_0} \tag{5.65}$$

$$\mathbf{B}_0=\frac{1}{N}\sum\nolimits_{i=1}^{N}\sum\nolimits_{j=1}^{m}\sum\nolimits_{k=1}^{m}\left.\frac{\partial\ln f_{ij}}{\partial\boldsymbol{\theta}}\right|_{\theta_0}\left.\frac{\partial\ln f_{ik}}{\partial\boldsymbol{\theta}'}\right|_{\theta_0}$$

的方差矩阵的三明治形式加以计算，其中，$f_{ij}=f(y_{ij}|\mathbf{x}_i,\boldsymbol{\theta})$。此外，与建立在联合密度基础上的 MLE 相比，偏 MLE 是无效的。有关进一步讨论，将在 6.9 节和 6.10 节给出。

5.8 非线性最小二乘法

NLS 估计量是把线性模型的 LS 估计自然推广到满足 $\mathrm{E}[y|\mathbf{x}]=g(\mathbf{x},\boldsymbol{\beta})$ 的非线性模型上，其中，$g(\cdot)$ 表示关于 $\boldsymbol{\beta}$ 为非线性的。本质上，分析与结果和线性最小二乘法的相同，其唯一的变化在于方差矩阵公式，回归元向量 $\mathbf{x}$ 被 $\partial g(\mathbf{x},\boldsymbol{\beta})/\partial\boldsymbol{\beta}|_{\hat{\beta}}$ 所代替，条件均值函数的导数在 $\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}$ 处计算。

对于微观经济计量分析来说，如同线性情况一样，必须要对异方差加以控制。通常，对异方差误差进行建模的估计量及推广，要比 MLSE 的有效性差，但它们在微观经济计量学中仍被广泛使用，这是因为它们依赖于比较弱的分布假设。

5.8.1 *非线性回归模型*

非线性回归模型(**nonlinear regression model**)定义纯量变量具有条件均值：

$$\mathrm{E}[y_i|\mathbf{x}_i]=g(\mathbf{x}_i,\boldsymbol{\beta}) \tag{5.66}$$

其中，$g(\cdot)$表示设定函数，$\mathbf{x}$ 表示解释变量的向量，而 $\boldsymbol{\beta}$ 表示 $K\times1$ 维参数向量。第 4 章的线性回归模型是 $g(\mathbf{x},\boldsymbol{\beta})=\mathbf{x}'\boldsymbol{\beta}$的特殊情况。

设定非线性函数 $\mathrm{E}[y|\mathbf{x}]$的普遍理由包括范围限制（例如，为了确保 $\mathrm{E}[y|\mathbf{x}]>0$）以及供给或需求的设定，或者源自生产者或消费者满足理论约束的成本或开支模型。一些广泛使用的非线性回归模型在表 5.5 中给出。

表 5.5　非线性最小二乘法：共同例子

模型	回归函数 $g(\mathbf{x},\boldsymbol{\beta})$
指数函数	$\exp\{(\beta_1x_1+\beta_2x_2+\beta_3x_3)\}$
自乘回归元形式	$\beta_1x_1+\beta_2x_2^{\beta_3}$
柯布—道格拉斯生产函数	$\beta_1x_1^{\beta_2}x_2^{\beta_3}$
CES 生产函数	$[\beta_1x_2^{\beta_2}+\beta_2x_2^{\beta_3}]^{1/\beta_3}$
非线性约束	$\beta_1x_1+\beta_2x_2+\beta_3x_3$，$-\beta_3=\beta_2\beta_1$

5.8.2 NLS 估计量

误差项被定义为因变量与其条件均值的差 $y_i-g(\mathbf{x}_i,\boldsymbol{\beta})$。非线性最小二乘法估计量是求残差平方和 $\sum_i(y_i-g(\mathbf{x}_i,\boldsymbol{\beta}))^2$ 的极小值，或者等价地求下式的极小值：

$$Q_N(\boldsymbol{\beta})=-\frac{1}{2N}\sum_{i=1}^{N}(y_i-g(\mathbf{x}_i,\boldsymbol{\beta}))^2 \tag{5.67}$$

其中，标度因子 1/2 简化了后面的分析。

若进行微分，则得到 NLS 一阶条件：

$$\frac{\partial Q_N(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}=\frac{1}{N}\sum_{i=1}^{N}\frac{\partial g_i}{\partial\boldsymbol{\beta}}(y_i-g_i)=\mathbf{0} \tag{5.68}$$

其中，$g_i=g(\mathbf{x}_i,\beta)$。这些条件把残差$(y-g)$限制成与$\partial g/\partial\boldsymbol{\beta}$是正交的，而不是同线性情况一样与 $\mathbf{x}$ 正交。$\hat{\boldsymbol{\beta}}_{\mathrm{NLS}}$不存在显式解，却可利用迭代方法进行计算（这一方法将由第 10 章给出）。

非线性回归模型能用矩阵记号以更简洁的方式表述，对观测值加以叠放整理，得到：

$$\begin{bmatrix}y_1\\ \vdots\\ y_N\end{bmatrix}=\begin{bmatrix}g_1\\ \vdots\\ g_N\end{bmatrix}+\begin{bmatrix}u_1\\ \vdots\\ u_N\end{bmatrix} \tag{5.69}$$

其中，$g_i=g(\mathbf{x}_i,\boldsymbol{\beta})$，或者等价地有：

$$\mathbf{y}=\mathbf{g}+\mathbf{u} \tag{5.70}$$

这里，$\mathbf{y}$、$\mathbf{g}$ 以及 $\mathbf{u}$ 均表示 $N\times1$ 维向量，它们的第 i 个元素分别为 y_i、g_i 以及 u_i。于是有：

$$Q_N(\boldsymbol{\beta})=-\frac{1}{2N}(\mathbf{y}-\mathbf{g})'(\mathbf{y}-\mathbf{g})$$

并且

$$\frac{\partial Q_N(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}=\frac{1}{N}\frac{\partial \mathbf{g}'}{\partial \boldsymbol{\beta}}(\mathbf{y}-\mathbf{g}) \tag{5.71}$$

其中：

$$\frac{\partial \mathbf{g}'}{\partial \boldsymbol{\beta}}=\begin{bmatrix} \frac{\partial g_1}{\partial \beta_1} & \cdots & \frac{\partial g_N}{\partial \beta_1} \\ \vdots & & \vdots \\ \frac{\partial g_1}{\partial \beta_K} & \cdots & \frac{\partial g_N}{\partial \beta_K} \end{bmatrix} \tag{5.72}$$

表示 $\mathbf{g}(\mathbf{x},\boldsymbol{\beta})'$对$\boldsymbol{\beta}$的 $K\times N$ 阶偏导数矩阵。

5.8.3 NLS 估计量的分布

NLS 估计量的分布将随 dgp 而变化。dgp 是能写成下式并具有可加性误差 u_i 的非线性回归：

$$y_i=g(\mathbf{x}_i,\boldsymbol{\beta}_0)+u_i \tag{5.73}$$

如果 dgp 中 $\mathrm{E}[y|\mathbf{x}]=g(\mathbf{x},\boldsymbol{\beta}_0)$，那么条件均值就被正确设定。从而，误差一定满足 $\mathrm{E}[u|\mathbf{x}]=0$。

给定 NLS 一阶条件(5.68)，一致性基本条件(5.25)变成：

$$\mathrm{E}[\partial g(\mathbf{x},\ \boldsymbol{\beta})/\partial\boldsymbol{\beta}|_{\beta_0}\times(y-g(\mathbf{x}_i,\boldsymbol{\beta}_0))]=\mathbf{0}$$

等价地讲，给定式(5.73)，我们要求 $\mathrm{E}[\partial g(\mathbf{x},\boldsymbol{\beta})/\partial\beta|_{\beta_0}\times u]=\mathbf{0}$。如果 $\mathrm{E}[u|\mathbf{x}]=0$，那么这个式子成立。因此，如同线性情况一样，一致性要求对条件均值的正确设定。然而，若 $\mathrm{E}[u|\mathbf{x}]\neq 0$，则实施一致估计，需要使用非线性工具方法(将在 6.5 节阐述)。

$\sqrt{N}(\hat{\boldsymbol{\beta}}_{\mathrm{NLS}}-\boldsymbol{\beta}_0)$的极限分布可利用一阶条件(5.68)的准确一阶泰勒级数展式来获得。对于位于$\hat{\boldsymbol{\beta}}_{\mathrm{NLS}}$与 $\boldsymbol{\beta}_0$ 之间的某个 $\boldsymbol{\beta}^+$。对式(5.18)中的 $\mathbf{A}_0$ 来说，得出：

$$\sqrt{N}(\boldsymbol{\beta}_{\mathrm{NLS}}-\boldsymbol{\beta}_0)=-\left(\frac{-1}{N}\sum_{i=1}^{N}\frac{\partial g_i}{\partial\boldsymbol{\beta}}\frac{\partial g_i}{\partial\boldsymbol{\beta}'}+\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2 g_i}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}(y_i-g_i)\bigg|_{\beta^+}\right)^{-1}\times\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\partial g_i}{\partial\boldsymbol{\beta}}u_i\bigg|_{\beta_0}$$

由于 $\mathrm{E}[u|\mathbf{x}]=0$，所以涉及$(\partial^2 g/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}')$的项被去掉，从而得以简化。因而，我们在渐近形式上只需要考虑：

$$\sqrt{N}(\boldsymbol{\beta}_{\mathrm{NLS}}-\boldsymbol{\beta}_0)=\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\partial g_i}{\partial\boldsymbol{\beta}}\frac{\partial g_i}{\partial\boldsymbol{\beta}'}\bigg|_{\beta_0}\right)^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\partial g_i}{\partial\boldsymbol{\beta}}u_i\bigg|_{\beta_0}$$

这与 OLS 的形式完全一样，只是 $\mathbf{x}_i$ 要用$\partial g_i/\partial\boldsymbol{\beta}'|_{\beta_0}$代替，参见 4.4.4 节。这就得到了下述命题，它类似于 OLS 估计量的命题 4.1。

命题 5.6(NLS 估计量的分布) 做出下述假设：

(i) 模型为式(5.73),也就是说,$y_i = g(\mathbf{x}_i, \boldsymbol{\beta}_0) + u_i$。

(ii) 在 dgp 中,$\mathrm{E}[u_i|\mathbf{x}_i]=0$ 且 $\mathrm{E}[\mathbf{uu}'|\mathbf{X}]=\boldsymbol{\Omega}_0$,其中,$\boldsymbol{\Omega}_{0,ij}=\sigma_{ij}$。

(iii) 均值函数 $g(\cdot)$ 满足 $g(\mathbf{x}, \boldsymbol{\beta}^{(1)}) = g(\mathbf{x}, \boldsymbol{\beta}^{(2)})$,当且仅当 $\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(2)}$。

(iv) 矩阵:

$$\mathbf{A}_0 = \operatorname{plim} \frac{1}{N} \sum_{i=1}^{N} \frac{\partial g_i}{\partial \boldsymbol{\beta}} \frac{\partial g_i}{\partial \boldsymbol{\beta}'}\bigg|_{\beta_0} = \operatorname{plim} \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \mathbf{g}'}{\partial \boldsymbol{\beta}} \frac{\partial \mathbf{g}}{\partial \boldsymbol{\beta}'}\bigg|_{\beta_0} \tag{5.74}$$

存在且是有限非奇异的。

(v) $N^{-1/2} \sum_{i=1}^{N} \partial g_i / \partial \boldsymbol{\beta} \times u_i |_{\beta_0} \xrightarrow{d} \mathcal{N}[0, \mathbf{B}_0]$,其中:

$$\mathbf{B}_0 = \operatorname{plim} \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \sigma_{ij} \frac{\partial g_i}{\partial \boldsymbol{\beta}} \frac{\partial g_j}{\partial \boldsymbol{\beta}'}\bigg|_{\beta_0} = \operatorname{plim} \frac{1}{N} \frac{\partial \mathbf{g}'}{\partial \boldsymbol{\beta}} \boldsymbol{\Omega}_0 \frac{\partial \mathbf{g}}{\partial \boldsymbol{\beta}'}\bigg|_{\beta_0} \tag{5.75}$$

那么,NLS 估计量 $\hat{\beta}_{\mathrm{NLS}}$ 被定义成一阶条件 $\partial N^{-1} Q_N(\beta) / \partial \boldsymbol{\beta} = \mathbf{0}$ 的根,它关于 β_0 是一致的,且满足:

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{\mathrm{NLS}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{A}_0^{-1} \mathbf{B}_0 \mathbf{A}_0^{-1}] \tag{5.76}$$

条件(i)~(iii)蕴含回归函数被正确地设定,而且回归元与误差项是不相关的,同时 $\boldsymbol{\beta}_0$ 是可识别的。误差可以是异方差性的,且对于不同的 i 是相关的。条件(iv)与(v)假定为了应用定理 53 而必须具备的限制结果。为使条件(v)得到满足,对不同 i 而言,需要在误差相关上施加一些约束。式(5.74)与式(5.75)中的关于 $\mathbf{X}$ 的概率极限是关于 dgp 的,如果 $\mathbf{X}$ 是非随机的,那么概率极限就是常规极限。

命题 5.6 中的矩阵 $\mathbf{A}_0$ 与 $\mathbf{B}_0$ 和 4.4.4 节中用$\partial g_i / \partial \boldsymbol{\beta}|_{\beta_0}$代替 $\mathbf{x}_i$ 的 OLS 估计量中的矩阵 $\mathbf{M}_{\mathbf{xx}}$与 $\mathbf{M}_{\mathbf{x\Omega x}}$一样。NLS 的渐近理论,与具有如此变化的 OLS 结果相同。

在球面误差下,$\boldsymbol{\Omega}_0 = \sigma_0^2 \mathbf{I}$,因而 $\mathbf{B}_0 = \sigma_0^2 \mathbf{A}_0$,且 $\mathrm{V}[\hat{\boldsymbol{\beta}}_{\mathrm{NLS}}] = \sigma_0^2 \mathbf{A}_0^{-1}$。于是,非线性最小二乘法在 LS 估计量中是渐近有效的。然而,横截面数据的误差不一定是异方差的。

给定命题 5.6,得到的 NLS 估计量的渐近分布表述为:

$$\hat{\boldsymbol{\beta}}_{\mathrm{NLS}} \overset{a}{\sim} \mathcal{N}[\boldsymbol{\beta}, (\mathbf{D}'\mathbf{D})^{-1} \mathbf{D}' \boldsymbol{\Omega}_0 \mathbf{D} (\mathbf{D}'\mathbf{D})^{-1}] \tag{5.77}$$

其中,导数矩阵 $\mathbf{D} = \partial \mathbf{g} / \partial \boldsymbol{\beta}'|_{\beta_0}$ 的第 i 行为$\partial g_i / \partial \boldsymbol{\beta}'|_{\beta_0}$[参见式(5.72)]。为了记号简单起见,不采用在 $\boldsymbol{\beta}_0$ 处的计算,同时我们假定可应用 LLN,因此,定义 $\mathbf{A}_0$ 与 $\mathbf{B}_0$ 中的 plim 算子可用 limE 来代替,然后省略 limit。后面几章将经常使用这种记号。

5.8.4 *NLS 的方差矩阵估计*

我们考察独立误差情形下常用的微观经济计量学的统计推断,其中,独立误差具有未知函数形式的异方差。这需要命题中曾定义的一致估计量。

因为 $\mathbf{A}_0$ 不涉及误差的矩,对于式(5.74)中已定义的 $\mathbf{A}_0$ 来说,可直接使用明显的估计量:

$$\hat{\mathbf{A}} = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial g_i}{\partial \boldsymbol{\beta}}\bigg|_{\hat{\beta}} \frac{\partial g_i}{\partial \boldsymbol{\beta}'}\bigg|_{\hat{\beta}} \tag{5.78}$$

给定对于不同 i 的独立性，由式(5.74)定义的 $\mathbf{B}_0$ 双重求和，被简化成单一求和：

$$\mathbf{B}_0 = \text{plim} \frac{1}{N}\sum_{i=1}^{N}\sigma_i^2 \frac{\partial g_i}{\partial \boldsymbol{\beta}} \frac{\partial g_i}{\partial \boldsymbol{\beta}'}\bigg|_{\beta_0}$$

就 OLS 估计量而论(参见 4.4.5 节)，只要求一致估计 $K\times K$ 阶矩阵和 $\mathbf{B}_0$。这并不要求 σ_0 的一致估计，即和式中的 N 个分量。

在怀特(White, 1980b)给出的条件下，下式：

$$\hat{\mathbf{B}} = \frac{1}{N}\sum_{i=1}^{N}\hat{u}_i^2 \frac{\partial g_i}{\partial \boldsymbol{\beta}} \frac{\partial g_i}{\partial \boldsymbol{\beta}'}\bigg|_{\hat{\beta}} = \frac{1}{N}\frac{\partial \mathbf{g}'}{\partial \boldsymbol{\beta}}\bigg|_{\hat{\beta}} \hat{\boldsymbol{\Omega}} \frac{\partial \mathbf{g}}{\partial \boldsymbol{\beta}'}\bigg|_{\hat{\beta}} \tag{5.79}$$

关于 $\mathbf{B}_0$ 是一致的，其中，$\hat{u}_i = y_i - g(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$，$\hat{\boldsymbol{\beta}}$关于 $\boldsymbol{\beta}_0$ 是一致的，同时：

$$\hat{\boldsymbol{\Omega}} = \text{Diag}[\hat{u}_i^2] \tag{5.80}$$

给出一些条件。这就导致下述 NLS 估计量的渐近方差矩阵的异方差性一致估计值：

$$\hat{\text{V}}[\hat{\boldsymbol{\beta}}_{\text{NLS}}] = (\hat{\mathbf{D}}'\hat{\mathbf{D}})^{-1}\hat{\mathbf{D}}'\hat{\boldsymbol{\Omega}}\hat{\mathbf{D}}(\hat{\mathbf{D}}'\hat{\mathbf{D}})^{-1} \tag{5.81}$$

其中，$\hat{\mathbf{D}} = \partial \mathbf{g}/\partial \boldsymbol{\beta}'|_{\hat{\beta}}$。该式与 4.4.5 节中的结果一样，只是要用 $\hat{\mathbf{D}}$ 代替回归元 $\mathbf{X}$。在实际应用中，可使用校正的自由度，因此，式(5.79)中的 $\hat{\mathbf{B}}$ 是用$(N-K)$去除，而不是用 N 去除。那么，式(5.81)的右边项应该用 $N/(N-K)$去乘。

对于不同 i，5.8.7 节将给出误差相关情况的推广。

5.8.5 指数回归例子

举一个事例，假定给定 $\mathbf{x}$ 时，y 具有指数条件均值，因而 $\text{E}[y|\mathbf{x}] = \exp(\mathbf{x}'\boldsymbol{\beta})$。此模型能表述成一个非线性回归：

$$y = \exp(\mathbf{x}'\boldsymbol{\beta}) + u$$

其中，误差项 u 满足 $\text{E}[u|\mathbf{x}] = 0$，并且误差是潜在异方差性的。

NLS 估计量具有一阶条件：

$$N^{-1}\sum\nolimits_i (y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta}))\exp(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i = \mathbf{0} \tag{5.82}$$

因此，$\hat{\boldsymbol{\beta}}_{\text{NLS}}$的一致性只要求条件均值被正确设定，满足 $\text{E}[y|\mathbf{x}] = \exp(\mathbf{x}'\boldsymbol{\beta}_0)$。这里，$\partial g/\partial \boldsymbol{\beta} = \exp(\mathbf{x}'\boldsymbol{\beta})\mathbf{x}$，所以一般 NLS 结果(5.81)会产生异方差性稳健的估计值。

$$\hat{\text{V}}[\hat{\boldsymbol{\beta}}_{\text{NLS}}] = \left(\sum\nolimits_i e^{2\mathbf{x}_i'\hat{\beta}}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sum\nolimits_i \hat{u}_i^2 e^{2\mathbf{x}_i'\hat{\beta}}\mathbf{x}_i\mathbf{x}_i'\left(\sum\nolimits_i e^{2\mathbf{x}_i'\hat{\beta}}\mathbf{x}_i\mathbf{x}_i'\right)^{-1} \tag{5.83}$$

其中，$\hat{u}_i = y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta}_{\text{NLS}})$。

5.8.6 加权 NLS 与 FGNLS

对于横截面数据来说，误差经常是异方差性的。于是，就控制异方差性而言，可行广义 NLS 比 NLS 更加有效。

与 ML 相比，可行广义非线性最小二乘法（FGNLS）通常还是差一些。一个著名的例外是，当 y 的条件密度是 LEF 密度时，FGNLS 渐近地等价于 MLE。一种特殊情况是，FGLS 渐近地等价于正态性下线性回归中的 MLE。

可行广义非线性最小二乘法

可行广义非线性最小二乘法估计量 $\hat{\beta}_{\text{FGNLS}}$ 使：

$$Q_N(\beta)=-\frac{1}{2N}(\mathbf{y}-\mathbf{g})'\boldsymbol{\Omega}(\hat{\gamma})^{-1}(\mathbf{y}-\mathbf{g}) \tag{5.84}$$

极大化，其中，假定 $\mathrm{E}[\mathbf{uu}'|\mathbf{x}]=\boldsymbol{\Omega}(\gamma_0)$，并且 $\hat{\gamma}$ 表示 γ_0 一致估计量。

如果对 NLS 估计量做出的假设得到满足，同时事实上 $\boldsymbol{\Omega}_0=\boldsymbol{\Omega}(\gamma_0)$，那么 FGNLS 估计量是一致的且渐近正态的，其估计渐近方差矩阵已由表 5.6 给出。方差矩阵估计值类似于线性 FGNL 的结果 $[\mathbf{X}'\boldsymbol{\Omega}(\hat{\gamma})^{-1}\mathbf{X}]^{-1}$，只是用 $\hat{\mathbf{D}}=\partial\mathbf{g}/\partial\beta'|_{\hat{\beta}}$ 来代替 $\mathbf{X}$。

表 5.6　非线性最小二乘法估计量与其渐近方差[a]

估计量	目标函数	估计渐近方差
NLS	$Q_N(\beta)=\frac{-1}{2N}\mathbf{u}'\mathbf{u}$	$(\hat{\mathbf{D}}'\hat{\mathbf{D}})^{-1}\hat{\mathbf{D}}'\hat{\boldsymbol{\Omega}}\hat{\mathbf{D}}(\hat{\mathbf{D}}'\hat{\mathbf{D}})^{-1}$
FGNLS	$Q_N(\beta)=\frac{-1}{2N}\mathbf{u}'\boldsymbol{\Omega}(\hat{\gamma})^{-1}\mathbf{u}$	$(\hat{\mathbf{D}}'\hat{\boldsymbol{\Omega}}^{-1}\hat{\mathbf{D}})^{-1}$
WNLS	$Q_N(\beta)=\frac{-1}{2N}\mathbf{u}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{u}$	$(\hat{\mathbf{D}}'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mathbf{D}})^{-1}\hat{\mathbf{D}}'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Omega}}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mathbf{D}}(\hat{\mathbf{D}}'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mathbf{D}})^{-1}$

[a] 函数是关于非线性回归模型的，其误差 $\mathbf{u}=\mathbf{y}-\mathbf{g}$ 已由式（5.70）和误差条件方差矩阵 Ω 定义。$\hat{\mathrm{D}}$ 表示关于 β 的条件均值向量导数在 β' 处的计算值。对于 GFNLS 来说，假定 $\hat{\mathrm{D}}$ 关于 Ω 是一致的。对于 NLS 与 WNLS 来说，异方差性稳健方差矩阵使用 $\hat{\Omega}$，而 $\hat{\Omega}$ 等于对角线为残差平方的对角矩阵，其估计值关于 Ω 不必是一致的。

FGNLS 估计量是求二次损失函数形式 $(\mathbf{y}-\mathbf{g})'\mathbf{V}(\mathbf{y}-\mathbf{g})$ 极小化估计量中最有效的一致估计量，其中，$\mathbf{V}$ 表示加权矩阵。

一般来讲，实施 FGNLS 需要 $N\times N$ 阶矩阵 $\boldsymbol{\Omega}(\hat{\gamma})$ 的形式。对很大的 N 来说，这在计算上是不可行的，但在实际应用中，通常 $\boldsymbol{\Omega}(\hat{\gamma})$ 具有对角结构，从而导致其逆具有解析解。

加权 NLS

若 $\boldsymbol{\Omega}_0$ 模型被错误设定，则尽管 FGNLS 方法是完全有效的，却会产生无效的标准误差估计。此外，我们考察介于 NLS 与 FGNLS 之间的一种方法，即对误差的方差矩阵模型加以设定，却获得稳健的标准误差。这种讨论反映在 4.5.2 节中。

加权非线性最小二乘法估计量 $\hat{\beta}_{\text{MNLS}}$ 使：

$$Q_N(\beta)=-\frac{1}{2N}(\mathbf{y}-\mathbf{g})'\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{y}-\mathbf{g}) \tag{5.85}$$

极大化，其中，$\boldsymbol{\Sigma}=\boldsymbol{\Sigma}(\gamma)$ 表示实用误差方差矩阵 $\hat{\boldsymbol{\Sigma}}=\boldsymbol{\Sigma}(\hat{\gamma})$，其中，$\hat{\gamma}$ 表示 γ 的估计值，而且在背离 FGNLS 情况下，有 $\boldsymbol{\Sigma}\neq\boldsymbol{\Omega}_0$。

在类似于对 *NLS* 估计量所做的那些假设下，同时假定 $\boldsymbol{\Sigma}_0=\text{plim}\,\hat{\boldsymbol{\Sigma}}$，WNLS 估计量是一致的且渐近正态的，其估计渐近方差矩阵已由表 5.6 给出。

这种估计量称为 WNLS，用以区分它与 FGNLS 的差别，这里假定 $\mathbf{\Sigma}=\mathbf{\Omega}_0$。就有效性而言，人们希望 WNLS 估计量位于 NLS 和 FGNLS 之间，如果误差方差矩阵模型选择不好，它的有效性就不如 NLS。NLS 与 OLS 估计量都是满足 $\mathbf{\Sigma}=\sigma^2\mathbf{I}$ 的 WNLS 特殊情况。

异方差性误差

异方差性的一个明显实用模型是 $\sigma_i^2=\mathrm{E}[u_i^2|\mathbf{x}_i]=\exp(\mathbf{z}_i'\boldsymbol{\gamma}_0)$，其中，向量 $\mathbf{z}$ 表示 $\mathbf{x}$ 的特定函数（例如，选取 $\mathbf{x}$ 的一些分量），并且利用指数形式确保正的方差。

于是，$\mathbf{\Sigma}=\mathrm{Diag}[\exp(\mathbf{z}_i'\boldsymbol{\gamma})]$，而 $\hat{\mathbf{\Sigma}}=\mathrm{Diag}[\exp(\mathbf{z}_i'\hat{\boldsymbol{\gamma}})]$，其中，$\hat{\boldsymbol{\gamma}}$ 可通过 NLS 残差平方和 $(y_i-g(\mathbf{x}_i,\hat{\boldsymbol{\beta}}_{\mathrm{NLS}}))^2$ 对 $\exp(\mathbf{z}_i'\hat{\boldsymbol{\gamma}})$ 的非线性回归来获得。由于 $\mathbf{\Sigma}$ 是对角的，$\mathbf{\Sigma}^{-1}=\mathrm{Diag}[1/\sigma_i^2]$。于是，式(5.84)可以简化，而 WNLS 估计量使

$$Q_N(\boldsymbol{\beta})=-\frac{1}{2N}\sum_{i=1}^{n}\frac{(y_i-g(\mathbf{x}_i,\boldsymbol{\beta}))^2}{\hat{\sigma}_i^2} \tag{5.86}$$

极大化。

由表 5.6 给出的 WNLS 估计量的方差矩阵导致：

$$\hat{\mathrm{V}}[\hat{\boldsymbol{\beta}}_{\mathrm{WNLS}}]=\left(\sum_{i=1}^{N}\frac{1}{\hat{\sigma}_i^2}\hat{\mathbf{d}}_i\hat{\mathbf{d}}_i'\right)^{-1}\left(\sum_{i=1}^{N}\hat{u}_i^2\frac{1}{\hat{\sigma}_i^4}\hat{\mathbf{d}}_i\hat{\mathbf{d}}_i'\right)\left(\sum_{i=1}^{N}\frac{1}{\hat{\sigma}_i^2}\hat{\mathbf{d}}_i\hat{\mathbf{d}}_i'\right)^{-1} \tag{5.87}$$

其中，$\hat{\mathbf{d}}_i=\partial g(\mathbf{x}_i,\boldsymbol{\beta})/\partial\boldsymbol{\beta}|_{\hat{\boldsymbol{\beta}}}$，而 $\hat{u}_i=y_i-g(\mathbf{x}_i,\hat{\boldsymbol{\beta}}_{\mathrm{WNLS}})$ 表示残差，在实际应用中，可使用修正的自由度，因而式(5.84)的右边可用 $N/(N-K)$ 去乘。倘若做出比较强的假设 $\mathbf{\Sigma}=\mathbf{\Omega}_0$，则 WNLS 变成 FGNLS，并且：

$$\hat{\mathrm{V}}[\hat{\boldsymbol{\beta}}_{\mathrm{FGNLS}}]=\left(\sum_{i=1}^{N}\frac{1}{\hat{\sigma}_i^2}\hat{\mathbf{d}}_i\hat{\mathbf{d}}_i'\right)^{-1} \tag{5.88}$$

可利用 NLS 程序计算 WNLS 与 FGNLS 估计量。首先，做 y_i 对 $g(\mathbf{x}_i,\boldsymbol{\beta})$ 的回归。其次，如果 $\sigma_i^2=\exp(\mathbf{z}_i'\boldsymbol{\gamma})$，那么通过 $(y_i-g(\mathbf{x}_i,\hat{\boldsymbol{\beta}}_{\mathrm{NLS}}))^2$ 对 $\exp(\mathbf{z}_i'\boldsymbol{\gamma})$ 的 NLS 回归来获得 $\hat{\boldsymbol{\gamma}}$。再次，实施 $y_i/\hat{\sigma}_i$ 对 $g(\mathbf{x}_i,\boldsymbol{\beta})/\hat{\sigma}_i$ 的 NLS 回归，$\hat{\sigma}_i^2=\exp(\mathbf{z}_i'\hat{\boldsymbol{\gamma}})$。这等价于对式(5.86)求极大值。源自这种变换回归的怀特稳健三明治标准误差给出了基于式(5.87)的稳健标准误差。通常，源自这种变换回归的非稳健标准误差，则给出基于式(5.88)的 FGNLS 标准误差。

对异方差性误差而言，一种非常引人注目的方法是，进一步探讨并利用 $\hat{\mathbf{\Omega}}=\mathrm{Diag}[\hat{u}_i^2]$ 来完成 FGNLS。然而，这将得到 $\boldsymbol{\beta}_0$ 的非一致参数估计值，因为 y_i 对 $g(\mathbf{x}_i,\boldsymbol{\beta})$ 的回归会简化成 $y_i/|\hat{u}_i|$ 对 $g(\mathbf{x}_i,\boldsymbol{\beta})/|\hat{u}_i|$ 的 FGNLS 回归。此技术因回归元与误差项之间相关的基本问题而受到损失。一些可供选择的半参数方法将在 9.7.6 节加以阐述，这些半参数方法不用对 $\mathbf{\Omega}_0$ 的函数形式进行设定，而是允许同可行 GLS 一样有效的估计量。

广义线性模型

实施加权 NLS 方法，需要对实用矩阵进行合理的设定。前面曾经阐述的特别方法是，令 $\hat{\sigma}_i^2=\exp(\mathbf{z}_i'\boldsymbol{\gamma})$，其中，$\mathbf{z}$ 通常表示 $\mathbf{x}$ 的子集。例如，在工资对受教育与其他控制变量的回归中，我们可以把异方差性更直接地建模成只有几个回归元的函数，回归元中最著名的是受教育程度。

某些类型的横截面数据，提供了使异方差性成为极为简约的正常模型。例如，对计数而言，泊松密度设定方差等于均值，因而 $\sigma_i^2=g(\mathbf{x}_i,\boldsymbol{\beta})$。这提供了异方差性的实用模型，与那些用于对条件均值进行建模的情况相比，这样做不必引入更多的参数。

这种把方差实用模型设为均值函数的方法，当然会出现在广义线性模型之中，这已经由 5.7.3 节和 5.7.4 节引进。由式(5.63)，建立在 LEF 密度基础上的准 MLE 的一阶条件具有下述形式：

$$\sum_{i=1}^{N}\frac{y_i-g(\mathbf{x}_i,\boldsymbol{\beta})}{\sigma_i^2}\times\frac{\partial g(\mathbf{x}_i,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}=\mathbf{0}$$

其中，假定 $\sigma_i^2=[c'(g(\mathbf{x}_i,\boldsymbol{\beta}))]^{-1}$对应于特殊 GLM 方差函数[参见式(5.60)]。例如，泊松分布、贝努利分布以及指数分布的 σ_i^2 分别等于 g_i、$g_i(1-g_i)$以及 $1/g_i^2$，其中，$g_i=g(\mathbf{x}_i,\boldsymbol{\beta})$。

在考虑 $\boldsymbol{\beta}$与$\hat{\sigma}_i^2$ 独立性的第一步中，就能求解这些一阶条件。在比较简单的两步方法中，给定 $\boldsymbol{\beta}$的初始 NLS 估计值，人们可计算$\hat{\sigma}_i^2=c'(g(\mathbf{x}_i,\hat{\boldsymbol{\beta}}))$，然后实施 $y_i/\hat{\sigma}_i$ 对 $g(\mathbf{x}_i,\boldsymbol{\beta})/\hat{\sigma}_i$ 的加权 NLS 回归。所得到的 $\boldsymbol{\beta}$估计量渐近地等价于直接求解准 MLE[参见古里耶克斯、蒙福特和特罗农(Gouriéroux, Monfort, and Trognan, 1984a)，或者卡梅伦和特里维迪(Cameron and Trivedi, 1986)]。因而，当密度是 LEF 密度时，FGNLS 渐近地等价于 ML 估计。为了预防对 σ_i^2 错误设定，推断建立在稳健三明治标准误差的基础上，或者令 $\hat{\sigma}_i^2=\hat{\alpha}[c'(g(\mathbf{x}_i,\hat{\boldsymbol{\beta}}))]^{-1}$，其中，$\hat{\alpha}$ 估计值已由 5.7.4 节给出。

5.8.7 *时间序列*

命题 5.6 中的一般 NLS 结果可应用于所有数据类型，包括时间序列数据。方差矩阵估计的后续结果是，关注于横截面的异方差性误差问题，但是，对它们很容易加以改进，以便适合具有序列相关误差的时间序列的问题。实际上，对时间序列情况利用谱方法的稳健方差矩阵估计结果，要优先于那些横截面的情况。

时间序列非线性回归模型是：

$$y_t=g(\mathbf{x}_t,\boldsymbol{\beta})+u_t,\quad t=1,\cdots,T$$

如果误差 u_t 是序列相关的，一种普遍做法是使用自回归移动平均(**autoregressive moving average**)或者 **ARMA**(p, q)模型：

$$u_t=\rho_1u_{t-1}+\cdots+\rho_pu_{t-p}+\varepsilon_t+\alpha_1\varepsilon_{t-1}+\cdots+\alpha_q\varepsilon_{t-q}$$

其中，ε_t 表示均值为 0 且方差为 σ^2 的 iid，ARMA 对模型参数施加约束，以确保平稳性和可逆性。ARMA 误差模型蕴含一种特殊结构，误差方差矩阵 $\boldsymbol{\Omega}_0=\boldsymbol{\Omega}(\boldsymbol{\rho},\boldsymbol{\alpha})$。

在时间序列情况下，ARMA 模型提供了 $\boldsymbol{\Omega}_0$ 的良好模型。与之相比，在横截面情况下，正确地对异方差性建模更为困难，这就导致对并不需要对 $\boldsymbol{\Omega}_0$ 模型进行设定的稳健推断进行更多的强调。

如果误差既是异方差性的又是序列相关的，会怎么样呢？如果误差是序列相关的，那么 NLS 估计量尽管是无效的，却是一致的，倘若 $\mathbf{x}_t$ 没有包括滞后因变量，

在此情况下，它变成非一致的。假定仅仅是譬如 l 个滞后的序列相关，怀特和多莫维茨(White and Domowitz, 1984)推广了式(5.79)，以便获得给定异方差性和未知函数形式的序列相关时，NLS 估计量的方差矩阵的稳健估计值。在实际应用中，使用归功于纽韦和韦斯特(Newey and West, 1987b)的稍微精炼的形式。这种精炼就是重新标度，以确保方差矩阵估计值是半正定的。人们还提出其他几种精炼，并且放松固定滞后长度的假设，因此，$l\rightarrow\infty$ 会以比 $N\rightarrow\infty$ 充分低的速率进行，这是可能的。这可以使误差有 AR 成分。

5.9 例子：ML 与 NLS 估计

极大似然估计与 NLS 估计、标准误差计算以及解释系数，均可以利用模拟数据来加以阐述。

5.9.1 *模型与估计量*

指数分布用于连续正的数据，譬如第 17 章将研究的著名持续期间数据。指数密度是指：

$$f(y)=\lambda e^{-\lambda y},\quad y>0,\ \lambda>0$$

其均值为 $1/\lambda$，方差为 $1/\lambda$。我们通过令：

$$\lambda=\exp(\mathbf{x}'\boldsymbol{\beta})$$

把回归元引入到此模型中，这确保 $\lambda>0$。注意到，这蕴含：

$$\mathrm{E}[y|\mathbf{x}]=\exp(-\mathbf{x}'\boldsymbol{\beta})$$

相反，一种可供选择的参数化设定 $\mathrm{E}[y|\mathbf{x}]=\exp(\mathbf{x}'\boldsymbol{\beta})$，因此，$\lambda=\exp(\mathbf{x}'\boldsymbol{\beta})$。注意，指数可以通过两种不同方式使用：用于密度与条件均值。

来自 y 对 $\mathbf{x}$ 回归的 OLS 估计量是非一致的，这是因为，当回归函数实际上是指数曲线时，它却拟合直线。

人们很容易获得 MLE。对数密度是 $\ln f(y|\mathbf{x})=\mathbf{x}'\boldsymbol{\beta}-y\exp(\mathbf{x}'\boldsymbol{\beta})$，从而 ML 一阶条件 $N^{-1}\sum_i(1-y_i\exp(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{x}_i=\mathbf{0}$，或者：

$$N^{-1}\sum_i\frac{y_i-\exp(-\mathbf{x}'\boldsymbol{\beta})}{\exp(-\mathbf{x}'\boldsymbol{\beta})}\mathbf{x}_i=\mathbf{0}$$

为了实施 NLS 回归，注意到，模型还能写成非线性回归：

$$y=\exp(-\mathbf{x}'\boldsymbol{\beta})+u$$

其中，误差项 u 具有 $\mathrm{E}[u|\mathbf{x}]=0$，尽管它是异方差的。这个模型指数条件均值的一阶条件，除了符号相反之外，已由式(5.82)给出，而且它显然导致了不同于 MLE 的估计量。

举一个加权 NLS 的例子，我们假定误差方差是与均值成比例的。于是，实用方差是 $\mathrm{V}[y]=\mathrm{E}[y]$，而加权最小二乘法可通过 $y_i/\hat{\sigma}_i$ 对 $\exp(-\mathbf{x}_i'\boldsymbol{\beta})/\hat{\sigma}_i$ 的 NLS 回归来实施，其中，估计量 $\hat{\sigma}_i^2=\exp(-\mathbf{x}_i'\hat{\boldsymbol{\beta}}_{\mathrm{NLS}})$ 的有效性不如 MLE，并且比 NLS 更

有效或不如 NLS 有效。

这里，由于我们知道 dgp，所以能实施可行广义 NLS。因为对于指数密度而言，$V[y]=1/\lambda^2$，所以方差等于均值平方，由此可得 $V[u|\mathbf{x}]=[\exp(-\mathbf{x}'\boldsymbol{\beta})]^2$。FGNLS 估计量通过 $\hat{\sigma}_i^2=[\exp(-\mathbf{x}_i'\hat{\boldsymbol{\beta}}_{\mathrm{NLS}})]^2$ 来估计 σ_i^2，并且能借助 $y_i/\hat{\sigma}_i$ 对 $\exp(-\mathbf{x}_i'\boldsymbol{\beta})/\hat{\sigma}_i$ 的 NLS 回归来实施。通常，FGNLS 的有效性不如 MLE 的有效性。在这个例子中，由于指数密度是 LEF 密度，所以它确实是完全有效的（参见 5.8.6 节末尾的讨论）。

5.9.2 模拟与结果

为了简单起见，我们考察一个截距与一个回归元的回归。数据生成过程是：

$$y|\mathbf{x}\sim\exp[\lambda]$$
$$\lambda=\exp(\beta_1+\beta_2 x)$$

其中，$x\sim\mathcal{N}[1,1^2]$且$(\beta_1,\beta_2)=(2,-1)$。为了最小化起因于抽样变量异性的估计差异，特别是标准误差，抽取一个样本量为 10 000 的大样本。就此处特定样本而言，y 的样本均值是 0.62，而 y 的样本标准差是 1.29。

表 5.7 给出 OLS、ML、NLS、WNLS 和 FGNLS 的估计值。还给出三种不同的标准误差估计值。默认回归产出得到非稳健的标准误差，这已在括号中已给出。对于 OLS 与 NLS 估计量来说，假定了 iid 误差，此处为不正确的假设；而对 MLE 来说，施加了 IM 等式，此处为有效假设，因为被假定的密度是 dgp。方括号中已给出的稳健标准误差使用稳健三明治方差估计 $N^{-1}\hat{\mathbf{A}}_{\mathrm{H}}^{-1}\hat{\mathbf{B}}_{\mathrm{OP}}\hat{\mathbf{A}}_{\mathrm{H}}^{-1}$，其中，$\hat{\mathbf{B}}_{\mathrm{OP}}$ 表示式(5.38)中给出的外积估计。这些估计值是异方差性一致的。对于 NLS 估计量的标准误差来说，一种可供选择的更好的估计已在大括号中给出（将在下一节进行解释）。此处，前面阐述的标准误差估计值在计算 $\hat{\mathbf{A}}$ 与 $\hat{\mathbf{B}}$ 时，使用了数值推导而不是解析推导。

表 5.7 指数事例：最小二乘法与估计[a]

变量	估计量				
	OLS	ML	NLS	WNLS	FGNLS
常数	−0.009 3	1.982 9	1.887 6	1.990 6	1.984 0
	(0.016 1)	(0.014 1)	(0.030 7)	(0.022 5)	(0.014 8)
	[0.017 2]	[0.014 4]	[0.142 1]	[0.035 9]	[0.014 6]
x	0.619 8	−0.989 6	−0.957 5	−0.996 1	−0.990 7
	(0.011 3)	(0.009 9)	(0.009 7)	(0.009 8)	(0.010 0)
	[0.025 4]	[0.009 9]	[0.061 2]	[0.022 4]	[0.010 1]
			{0.088 0}		
Ln L	—	−208.71	−232.98	−208.93	−208.72
R^2	0.232 6	0.390 6	0.391 3	0.390 2	0.390 6

a 除了 OLS 之外，所有估计量都是一致的。三种可供选择的标准误差估计值都已给出，小括号中为非稳健估计值，方括号中为稳健外积的估计值，而大括号中为可供选择的稳健估计值。条件是指数分布的，其截距为 2 且斜率为 −1。样本量为 $N=10\ 000$。

5.9.3 估计值与标准误差的比较

OLS 估计量是非一致的，得出的估计值在指数 dgp 下与(β_1, β_2)不相关。

剩下的估计量都是一致的，而 ML、NLS、WNLS 以及 FGNLS 估计量，都位于真实参数值(2, −1)的两个标准误差之内，其中，对于 NLS 来说，必须使用稳健标准误差。FGNLS 估计值十分接近 ML 估计值，即在 LEF 中利用 dgp 的结果。

对 MLE 而言，非稳健标准误差与稳健标准误差都非常相似。正如人们所希望的，它们是渐近等价的(因为如果 MLE 是建立在真实密度的基础上，信息矩阵等式就成立)。而此时的样本量是很大的。

对 NLS 而言，非稳健的标准误差是无效的，因为 dgp 具有异方差性误差，并且高估了 NLS 估计值的准确性。NLS 的稳健方差矩阵估计值的公式已由式(5.81)给出，其中，$\hat{\mathbf{\Omega}}=\mathrm{Diag}[\hat{u}_i^2]$。使用 $\hat{\mathbf{\Omega}}=\mathrm{Diag}[\hat{\mathrm{E}}[u_i^2]]$的一种可供选择的方法由大括号给出，其中，$\hat{\mathrm{E}}[u_i^2]=[\exp(-\mathbf{x}_i'\hat{\boldsymbol{\beta}})]^2$。对斜率函数而言，两个估计值 0.061 2 与 0.088 0 确实不同。因为$\hat{u}_i^2=(y_i-\exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}}))^2$ 不同于 $\exp[(-\mathbf{x}_i'\hat{\boldsymbol{\beta}})]^2$，所以才出现这种差异。更一般地讲，甚至在相当大的样本中，利用外积估计的标准误差是有偏的。NLS 的有效性相当不如 MLE 的有效性。其标准误差为利用大括号中更可取的估计值 MLE 的那些标准误差的许多倍。

WNLS 估计量没有使用异方差性的正确模型，因此，其非稳健的标准误差与稳健的标准误差再次出现不同。一旦利用稳健的标准误差，WNLS 估计量就比 NLS 估计量更有效，但不如 MLE 估计量有效。

在这个事例中，已知位于 LEF 中的 dgp 的结果估计量，FGNLS 与 MLE 估计量是一样有效的。此结果表明，FGNLS 的系数及标准误差非常接近于 MLE 的情况。对 FGNLS 估计量而言，正如人们所料，稳健标准误差与非稳健标准误差本质上是相同的，因为它正确设定了异方差性模型。

表 5.7 还报告出估计对数似然，$\ln L=\sum_i[\mathbf{x}_i'\hat{\boldsymbol{\beta}}-\exp(-\mathbf{x}_i'\hat{\boldsymbol{\beta}})y_i]$，而 $R^2=1-\sum_i(y_i-\hat{y}_i)^2/\sum_i(y_i-\bar{y})^2$ 测量在 ML、NLS、WNLS 以及 FGNLS 估计值处的计算值，其中，$\hat{y}_i=\exp(-\mathbf{x}_i'\hat{\boldsymbol{\beta}})$。各种模型的 R^2 略有不同，对 NLS 估计量来说是最小的，如同人们期望的，这是因为 NLS 对$\sum_i(y_i-\hat{y}_i)^2$ 求极小化。正如人们所料，通过 MLE 求对数似然极大值，对 NLS 估计量而言是相当小的。

5.9.4 解释系数

关注内容在于，当 x 变化时所引起的 $\mathrm{E}[y|x]$变动。我们考察由表 5.7 给出 $\hat{\beta}_2=-0.99$ ML 的估计值。

条件均值 $\exp(-\beta_1-\beta_2 x)$是单指标形式，因此，如果具有系数 β_3 的另外回归元 z 被包括进来，那么 z 变化 1 个单位的边际效应是 x 变化 1 个单位所引起边际效应的 $\hat{\beta}_3/\hat{\beta}_2$ 倍(参见 5.2.4 节)。

条件均值关于 x 是单调递减的，因而 $\hat{\beta}_2$ 的符号与边际效应的符号相反(参见 5.2.4 节)。这里，x 增大的边际效应引起条件均值增大，这是因为 $\hat{\beta}_2$ 是负的。

现在，我们利用微分法考察 x 变化引起的边际效应。这里，$\partial\mathrm{E}[y|x]/\partial x=$

$-\beta_2\exp(-\mathbf{x}'\boldsymbol{\beta})$随着点 x 计算值变化而变动，在样本中其范围从 0.01～19.09。样本平均响应是 $0.99\ N^{-1}\sum_i\exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}})=0.61$。响应在样本均值处的计算值为 $0.99\times\exp(\bar{\mathbf{x}}'\hat{\boldsymbol{\beta}})=0.37$，这显得相当小。由于$\partial\mathrm{E}[y|x]/\partial x=-\beta_2\mathrm{E}[y|x]$，边际效应的另外一个估计值是 $0.99\bar{y}=0.61$。

有限差分法会产生不同的估计边际效应。当 $\Delta x=1$ 时，我们可得到 $\Delta\mathrm{E}[y|\mathbf{x}]=(e^{\beta_2}-1)\exp(-\mathbf{x}'\boldsymbol{\beta})$（参见 5.2.4 节）。这得出样本的平均响应是 1.04，而不是 0.61。然而，如果 Δx 很小，有限差分法与微分法是一致的。

前面的边际效应是可加的。对于指数条件均值来说，我们还可以考察乘法或比例的边际效应（参见 5.2.4 节）。例如，x 变化 0.1 个单位，会预测 $\mathrm{E}[y|x]$的增大比例为 0.1×0.99 或增大 9.9%。同理，有限差分法将会产生不同的估计值。

这些测量中，哪一个是最有用的呢？对单指标形式加以约束是非常有用的，这是因为回归元的相对影响能够被立刻计算出来。对响应数值来说，最准确的是，利用非微分法，计算出回归元变化 c 个单位时样本的平均响应，其中，数值 c 是一个有意义的数量，比如 x 变化一个标准差。

对于 NLS、WNKS 和 FGNLS 估计值来说，可进行类似计算，得出相似结果。对 OLS 估计量而言，注意到，x 系数能被解释为 x 变化时样本平均边际效应（参见 4.7.2 节）。这里，OLS 估计值 $\hat{\beta}_2=0.61$ 与前面利用指数 MLE 计算出的样本平均响应的两位小数值相同。这里的 OLS 提供了样本平均边际效应的良好估计值，尽管对 x 的任何特殊值而言，它提供了边际响应非常不好的估计值。

5.10 应用研究

为了获得 5.6.1 节引入的标准模型极大似然估计量，大多数经济计量学软件包都提供简单的命令。对其他密度而言，许多软件包都提供 ML 程序，为用户配备了密度方程以及可能的一阶导数甚至二阶导数。类似地，就 NLS 而言，软件包给出 NLS 程序的条件均值方程。对于一些非线性模型及数据集来说，软件包中配备的 ML 与 NLS 程序在求估计值时会遇到计算上的困难。在这种背景下，有必要使用作为外接式附件 GAUSS、Matlab 和 OX 的更稳健的最优化程序。GAUSS、Matlab 和 OX 是非线性建模的较好工具，但要投入较大的初始学习成本。

对于横截面数据来说，使用基于方差矩阵的三明治形式标准误差已成为标准方法。这些是经常提供的命令选项。就 NL 估计量而言，这给出异方差性一致的标准误差。对极大似然而言，人们应该认识到，除了需要使用三明治误差以外，对密度的错误设定会导致非一致性。

通常，不能直接对非线性模型的参数给予解释，并且，一种好的做法是另外计算回归元上的变动引起的隐含边际效应（参见 5.2.4 节）。有些软件包会自动进行这种计算，对其他几种后估计方法来说，需要利用已保存的回归系数进行编码。

5.11 文献注释

关于极值估计量的渐近理论研究成果，纽韦和麦克法登（Newey and McFad-

den，1994，第 2 115 页）给出了简略历史。雨宫（Amemiya，1973）对重要的经济计量学进展给出评价，雨宫发展了可用于 Tobit 模型 MLE 的一般定理。有益的教科书式长篇评论，包括由加伦特（Gallant，1987）、加伦特和怀特（Gallant and White，1987）、比勒斯（Berens，1993）以及怀特（White，1994，2001a）撰写的著作。

许多书都曾给出统计基础，包括雨宫（Amemiya，1985，第 3 章）、戴维森和麦金农（Davidson and MacKinnon，1993，第 4 章）、格林（Greene，2003，附录 D）、戴维森（Davidson，1994）以及扎曼（Zaman，1996）。

5.3 一般极值估计结果的表述大量地利用雨宫（Amemiya，1985，第 4 章）的成果，但在扩展程度上远不如纽韦和麦克法登（Newey and McFadden，1994）。后者的参考书是非常综合的。

5.4 估计方程用于广义线性模型文献之中[参见麦卡拉和内尔德（McCullagh and Nelder，1989）]。经济计量学家把这些内容归入广义矩方法中（参见第 6 章）。

5.5 第 7 章将详细阐述统计推断。

5.6 ML 估计的一般结果，参见费希尔的开创性文章（Fisher，1992），包括有效性和似然法与反概率或贝叶斯方法、矩方法估计的比较。

5.7 现代应用中，经常使用准 ML 框架以及方差矩阵的三明治估计[参见怀特（White，1982，1994）]。在统计学中，此方法称为广义线性模型，麦卡拉和内尔德（McCullagh and Nelder，1989）的书已成为标准参考书。

5.8 类似于 NLS 估计，方差矩阵的三明治估计值用于需要相对弱的假设的误差过程中。在经济计量学中，由怀特（White，1980a，c）撰写的论文对统计推断产生了重大影响。渐近理论的推广与详细评述则由怀特和多莫维茨（White and Domowitz，1984）给出。雨宫（Amemiya，1983）对非线性回归进行了全面评述。

习 题

5-1 假定我们得到可以产生预测条件均值的模型估计值 $\hat{\mathrm{E}}[y|x]=\exp(1+0.01x)/[1+\exp(1+0.01x)]$。假定有容量为 100 的样本，取值为整数值 1，2，…，100。求下述估计边际效应$\partial\hat{\mathrm{E}}[y|x]/\partial x$ 的估计值。

(a) 所有观测值的平均边际效应。

(b) 平均观测值的边际效应。

(c) 当 $x=90$ 时的边际效应。

(d) 利用有限差分法计算，当 $x=90$ 时变化一个单位的边际效应。

5-2 考察下述伽玛分布的特殊单一参数情况，$f(y)=(y/\lambda^2)\exp(-y/\lambda)$，$y>0$，$\lambda>0$。对这个分布而言，可以证明，$\mathrm{E}[y]=2\lambda$ 且 $\mathrm{V}[y]=2\lambda^2$。此处，我们引入回归元，并假定在真实模型中，参数 λ 依照 $\lambda_j=\exp(\mathbf{x}_i'\boldsymbol{\beta})/2$ 而依赖于回归元。因而，$\mathrm{E}[y_i|\mathbf{x}_i]=\exp(\mathbf{x}_i'\boldsymbol{\beta})$且 $\mathrm{V}[y_i|\mathbf{x}_i]=[\exp(\mathbf{x}_i'\boldsymbol{\beta})]^2/2$。假设对于不同 i，数据是独立的且 $\mathbf{x}_i$ 是非随机的。而且在 dgp 中，有 $\boldsymbol{\beta}=\boldsymbol{\beta}_0$。

(a) 证明，此伽玛模型的对数似然函数（由 N^{-1}来标度）是 $Q_N(\boldsymbol{\beta})=N^{-1}\sum_i\times\{\ln y_i-2\mathbf{x}_i'\boldsymbol{\beta}+2\ln 2-2y_i\exp(-\mathbf{x}_i'\boldsymbol{\beta})\}$。

(b) 求 $\text{plim}\, Q_N(\boldsymbol{\beta})$。你可以假定，为了使用任何 LLN，所需的假设都得以满足。（提示：$\mathrm{E}[\ln y_i]$依赖于 $\boldsymbol{\beta}_0$ 但不依赖于 $\boldsymbol{\beta}$。）

(c) 证明，作为 $Q_N(\boldsymbol{\beta})$的局部极大值的 $\hat{\boldsymbol{\beta}}$，关于 $\boldsymbol{\beta}_0$ 是一致的。叙述做出的任何假设。

(d) 现在，阐述为了验证(b)部分，你会使用什么样的 LLN。并且，为了应用这个定律，需要什么额外信息；如果有的话。请做出简要回答。这里不要求给出正式证明。

5-3 继续习题 5-2 中的伽玛模型。

(a) 证明：$\partial Q_N(\boldsymbol{\beta})/\partial\boldsymbol{\beta}=N^{-1}\sum_i 2[(y_i-\exp(\mathbf{x}_i'\boldsymbol{\beta}))/\exp(\mathbf{x}_i'\boldsymbol{\beta}))]\mathbf{x}_i$。

(b) 为使 $\hat{\boldsymbol{\beta}}$成为一致的，由一阶条件预示的什么根本条件必须得到满足？

(c) 应用中心极限定理，求 $\sqrt{N}\partial Q_N/\partial\boldsymbol{\beta}|_{\boldsymbol{\beta}_0}$ 的极限分布。此处，你能够假定中心极限定理所需的假设都得到满足。

(d) 为了验证(c)部分，叙述你会使用什么样的 CLT。并且，为了应用这个定律，需要什么额外信息；如果有的话。请做出简要回答。这里不要求给出正式证明。

(e) 求$\partial^2 Q_N/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'|_{\boldsymbol{\beta}_0}$的概率极限。

(f) 结合前面结果，求$\sqrt{N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)$的极限分布。

(g) 已知(f)部分，阐述如何在水平 0.05 上检验 H_0：$\beta_{0j}\geqslant\beta_j^*$ 对 H_a：$\beta_{0j}<\beta_j^*$，其中，β_j 表示 $\boldsymbol{\beta}$的第 j 个分量。

5-4 非负整数变量 y 服从几何分布，具有密度（或者更正式地为概率质量函数）$f(y)=(y+1)(2\lambda)^y(1+\lambda)^{-(y+0.5)}$，$y=0,1,2,\cdots,\lambda>0$。于是，$\mathrm{E}[y]=\lambda$ 且 $\mathrm{V}[y]=\lambda(1+2\lambda)$。引入回归元，并假定 $\gamma_i=\exp(\mathbf{x}_i'\boldsymbol{\beta})$。假定对于不同 i，数据是独立的，并且 $\mathbf{x}_i$ 是非随机的，并且在 dgp 中 $\boldsymbol{\beta}=\boldsymbol{\beta}_0$。

(a) 对该模型重复习题 5-2 中的问题。

(b) 对该模型重复习题 5-3 中的问题。

5-5 假定从一个样本得出估计值 $\hat{\theta}_1=5,\hat{\theta}_2=3,\mathrm{se}[\hat{\theta}_1]=2$，而 $\mathrm{se}[\hat{\theta}_2]=1$。同时，$\hat{\theta}_1$ 与 $\hat{\theta}_2$ 之间的相关系数等于 0.5。一旦假定参数估计值具有渐近正态性，执行下述水平为 0.05 的检验。

(a) 检验 H_0：$\theta_1=0$ 对 H_a：$\theta_1\neq 0$。

(b) 检验 H_0：$\theta_1=2\theta_2$对 H_a：$\theta_1\neq 2\theta_2$。

(c) 检验 H_0：$\theta_1=0$，$\theta_2=0$ 对 H_a：θ_1 和 θ_2 中至少有一个不为 0。

5-6 考察非线性回归模型 $y=\exp(\mathbf{x}'\boldsymbol{\beta})/[1+\exp(\mathbf{x}'\boldsymbol{\beta})]+u$，其中，误差项可能是异方差的。

(a) 这种限制 $\mathrm{E}[y|\mathbf{x}]$会位于什么范围内？

(b) 给出 NLS 估计量的一阶条件。

(c) 利用结果(5.77)，求 NLS 估计量的渐近分布。

5-7 这个问题假定可以使用软件计算 NLS 与 ML 估计。考察习题 5-2 的伽玛回归模型。一种合适的伽玛变量能由 $y=-\lambda\ln r_1-\lambda\ln r_2$ 生成，其中，$\lambda=\exp(\mathbf{x}'\boldsymbol{\beta})/2$，而 r_1 与 r_2 是从均匀分布[0，1]中随机抽取的。令 $\mathbf{x}'\boldsymbol{\beta}=\beta_1+\beta_2 x$。当

$\beta_1=-1.0$，$\beta_2=1$，以及 $\mathbf{x}\sim\mathcal{N}[0, 1]$时，生成一个容量为 10 000 的样本。

(a) 求 y 对 $\exp(\beta_1+\beta_2 x)$的 NLS 回归的 β_1 与 β_2 估计值。

(b) 这里应该使用三明治标准误差吗？

(c) 求 y 对 $\exp(\beta_1+\beta_2 x)$的 NLS 回归的 β_1 与 β_2 ML 估计值。

(d) 这里应该使用三明治标准误差吗？

6 广义矩方法与系统估计

6.1 引 论

上一章关注 m 估计，包括 ML 与 NLS 估计。现在，我们考察一类更广泛的极值估计量，即建立在矩方法(MM)与广义矩方法(GMM)基础上的估计量。

矩方法与广义矩方法的基础是对总体矩条件的集合进行设定，而总体矩条件涉及数据与未知参数。矩方法估计量求解相应总体条件的样本矩条件。例如，样本均值是总体均值的矩方法估计量。在一些情况下，对于矩方法估计量来说，可能不存在明显解析解，但对其求解数值解还是可行的。于是，该估计量就是 5.4 节曾简略介绍的估计方程估计量的一个例子。

然而，在一些情况下，矩方法估计或许是行不通的，因为存在着比参数还多的矩条件和待求解方程。一个重要例子就是过度识别模型中的工具变量估计。归功于汉森(Hansen，1982)的广义矩方法估计量扩展了矩方法，以便适应这种情况。

广义矩方法估计量定义一类估计量，利用各种不同的总体矩条件，可获得不同的广义矩方法估计量，正如不同的设定密度会产生不同的 ML 估计量一样。甚至当可能有可供选择的表示时，我们仍强调基于矩的估计方法，这是因为它提供了一种统一的估计方法，并且提供一种从线性到非线性模型扩展方法的明确途径。

广义矩方法估计的基础由 6.2 节和 6.3 节给出，这两节分别阐述统计推断的解释性例子和渐近结果。而本章其余部分详述更专门化的估计量。6.4 节与 6.5 节阐述工具变量估计量。对线性模型而言，4.8 节和 4.9 节的研究或许是充分的，但对非线性模型的扩展来说，则要使用广义矩方法。6.6 节涵盖计算时序两步 m 估计量的标准误差的方法。6.7 节与 6.8 节阐述最小距离估计量、广义矩方法的变形，以及经验似然估计量，即针对广义矩方法的可供选择的估计量。在微观经济计量研究中，相对而言，仅有很小一部分所使用的系统估计方法将在 6.9 节与 6.10 节加以讨论。

本章从广义矩方法观点出发，对许多估计方法重新考察。利用这些方法和实际数据进行应用研究，包括对 4.9.6 节中线性工具变量(IV)的应用，以及对 22.3 节中线性面板广义矩方法的应用。

6.2 例 子

广义矩方法估计量的建立基础是，总体矩条件导致能用于估计参数样本矩条件的类比原理(参见5.4.2节)。本节将提供关于这个原理的几个重要应用，所得到的估计量的一些性质可参考6.3节。

6.2.1 线性回归

当 y 是iid的且均值为 μ 时，矩方法(**method of moments**)的经典例子是对总体均值的估计。总体中有：

$$\mathrm{E}[y-\mu]=0$$

通过用样本的平均算子 $N^{-1}\sum_{i=1}^{N}(\cdot)$ 代替总体期望算子 $E[\cdot]$，就会得到相应样本矩：

$$\frac{1}{N}\sum_{i=1}^{N}(y_i-\mu)=0$$

然后求解 μ，得出估计量 $\hat{\mu}_{\mathrm{MM}}=N^{-1}\sum_i y_i=\bar{y}$。总体均值的矩方法估计正是样本均值。

这种方法能被推广到线性回归模型 $y=\mathbf{x}'\boldsymbol{\beta}+u$ 上，其中，$\mathbf{x}$ 与 $\boldsymbol{\beta}$ 都表示 $K\times1$ 维向量。假定误差项 u 具有以回归元为条件的零均值。单个条件约束 $\mathrm{E}[u|\mathbf{x}]=0$ 会产生 K 个无条件的条件矩 $\mathrm{E}[\mathbf{x}u]=\mathbf{0}$，这是因为：

$$\mathrm{E}[\mathbf{x}u]=\mathrm{E}_{\mathbf{x}}[\mathrm{E}[\mathbf{x}u|\mathbf{x}]]=\mathrm{E}_{\mathbf{x}}[\mathbf{x}\mathrm{E}[u|\mathbf{x}]]=\mathrm{E}_{\mathbf{x}}[\mathbf{x}\cdot 0]=\mathbf{0} \tag{6.1}$$

推导中利用了期望迭代定律[1](**law of iterated expectations**)(参见A.8节)以及 $\mathrm{E}[u|\mathbf{x}]=0$ 的假设。因而，若误差具有条件零均值，则：

$$\mathrm{E}[\mathbf{x}(y-\mathbf{x}'\boldsymbol{\beta})]=\mathbf{0}$$

矩方法估计量是相应样本矩条件

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i(y_i-\mathbf{x}_i'\boldsymbol{\beta})=\mathbf{0}$$

的解。这就得到，$\hat{\boldsymbol{\beta}}_{\mathrm{MM}}=(\sum_i\mathbf{x}_i\mathbf{x}_i')^{-1}\sum_i\mathbf{x}_i y_i$。

因此，OLS估计量是矩方法估计的一种特殊情况。不过，OLS估计量的矩方法推导显著地不同于通常对残差平方和求极小值的推导。

6.2.2 非线性回归

对于非线性回归而言，若回归误差是可加的，则矩方法就简化成NLS。对于更一般的具有非可加误差(下面将定义)的非线性回归来说，矩方法将会得出一致估计量，而NLS却是非一致的。

---

〔1〕 又称为重期望定律。——译者注

由 5.8.3 节，具有可加误差(**additive error**)的非线性回归模型，就是设定：

$$y=g(\mathbf{x},\boldsymbol{\beta})+u$$

的模型。类似于线性模型情况，矩方法可得到 $\mathrm{E}[u|\mathbf{x}]=0$，这蕴含着 $y-g(x-\boldsymbol{\beta})=\mathbf{0}$，其中，$\mathbf{h}(\mathbf{x})$表示 $\mathbf{x}$ 的任意函数。从 6.3.7 节的内容出发，对 $\mathbf{h}(\mathbf{x})=\partial g(\mathbf{x},\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ 进行特殊选择，则会产生相应的样本矩条件，而这些样本矩条件等于由 5.8.2 节给出的 NLS 估计量的一阶条件。

具有非可加误差(**nonadditive error**)的更一般的回归模型是设定：

$$u=r(y,\mathbf{x},\boldsymbol{\beta})$$

其中，再次有 $\mathrm{E}[u|\mathbf{x}]=0$，但 y 不再被约束为 u 的可加函数。例如，在泊松回归中，人们可以定义标准化误差 $u=[y-\exp(\mathbf{x}'\boldsymbol{\beta})]/[\exp(\mathbf{x}'\boldsymbol{\beta})]^{1/2}$，由于 y 具有等于 $\exp(\mathbf{x}'\boldsymbol{\beta})$的条件均值与条件方差，所以 $\mathrm{E}[u|\mathbf{x}]=0$ 且 $\mathrm{V}[u|\mathbf{x}]=1$。

已知非可加误差，NLS 估计量是非一致的。对 $N^{-1}\sum_i u_i^2=N^{-1}\sum_i r(y_i,\mathbf{x}_i,\boldsymbol{\beta})^2$ 求极小值，得出一阶条件：

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial r(y_i,\ \mathbf{x}_i,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}r(y_i,\ \mathbf{x}_i,\boldsymbol{\beta})=\mathbf{0}$$

这里，y_i 在乘积的两项中都出现了，即使 $\mathrm{E}[r(\cdot)|\mathbf{x}]=0$，也无法保证这个乘积具有零期望。对可加误差 $r(\cdot)=y-g(\mathbf{x},\boldsymbol{\beta})$而言，这种非一致性就不会产生，因为 $\partial r(\cdot)/\partial\boldsymbol{\beta}=-\partial g(\mathbf{x},\boldsymbol{\beta})/\partial\boldsymbol{\beta}$，所以仅有乘积中的第二项依赖于 y。

基于矩方法，会产生一致估计量。$\mathrm{E}[u|\mathbf{x}]=0$ 的假设蕴含着：

$$\mathrm{E}[\mathbf{h}(\mathbf{x})r(y,\mathbf{x},\boldsymbol{\beta})]=\mathbf{0}$$

其中，$\mathbf{h}(\mathbf{x})$表示 $\mathbf{x}$ 的函数。若 $\dim[\mathbf{h}(\mathbf{x})]=K$，则利用相应样本矩：

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{h}(\mathbf{x}_i)r(y_i,\mathbf{x}_i,\boldsymbol{\beta})=\mathbf{0}$$

得到 $\boldsymbol{\beta}$的一致估计量，其解可通过数值方法求出。

6.2.3 *极大似然法*

库尔贝克—莱布勒(Kullback-Leibler)信息准则已在 5.2.2 节定义。由此定义，若 $\mathrm{E}[\mathbf{s}(\boldsymbol{\theta})]=\mathbf{0}$，其中，$\mathbf{s}(\boldsymbol{\theta})=\partial\ln f(y|\mathbf{x},\boldsymbol{\theta})/\partial\boldsymbol{\theta}$，而 $f(y|\mathbf{x},\boldsymbol{\theta})$表示条件密度，则就会出现库尔贝克—莱布勒信息准则(KLIC)的局部极大值。

若用样本矩代替总体矩，则会得到作为 $N^{-1}\sum_i\mathbf{s}_i(\boldsymbol{\theta})=\mathbf{0}$ 解的估计量 $\hat{\boldsymbol{\theta}}$ 。由于 ML 的一阶条件存在，所以 MLE 可以成为 MM 估计量。

6.2.4 *额外矩约束*

如果矩条件比要估计的参数多一些，那么利用额外矩可改进估计有效性，但要采用正规矩方法。

无效估计量的一个简单例子是样本均值。这是总体均值的无效估计量，除非数据是出自正态分布或者指数分布族的某些其他分布的随机样本。一种改进有效

性的方法就是使用一种可供选择的估计量。假如分布是对称的，关于 μ 为一致的样本中位数就会更为有效。很明显，若分布是完全设定的，就能使用 MLE，然而，我们这里反而通过利用额外矩约束改进有效性。

考察线性回归模型中关于 $\boldsymbol{\beta}$ 的估计。甚至在假定同方差误差下，OLS 估计量仍是无效的，除非误差服从正态分布。由 6.2.1 节知道，OLS 估计量是建立在 $\mathrm{E}[\mathbf{x}u]=\mathbf{0}$ 基础上的 MM 估计量。现在，做出另外的矩假设：误差是条件对称的，所以 $\mathrm{E}[u^3|\mathbf{x}]=0$，从而 $\mathrm{E}[\mathbf{x}u^3]=\mathbf{0}$。于是，对 $\boldsymbol{\beta}$ 的估计可建立在 $2K$ 个矩条件：

$$\begin{bmatrix}\mathrm{E}[\mathbf{x}(y-\mathbf{x}'\boldsymbol{\beta})]\\ \mathrm{E}[\mathbf{x}(y-\mathbf{x}'\boldsymbol{\beta})^3]\end{bmatrix}=\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}$$

的基础上。MM 估计量试图希望估计 $\boldsymbol{\beta}$ 作为相应样本矩条件 $N^{-1}\sum_i\mathbf{x}_i(y_i-\mathbf{x}_i'\boldsymbol{\beta})=\mathbf{0}$ 与 $N^{-1}\sum_i\mathbf{x}_i(y_i-\mathbf{x}_i'\boldsymbol{\beta})^3=\mathbf{0}$ 的解。然而，对 $2K$ 个方程与只有 K 个未知参数 $\boldsymbol{\beta}$ 来说，满足所有这些样本矩条件是不可能的。

相反，广义矩方法估计量利用二次损失，尽可能地使样本矩接近于 0。那么，$\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}$ 对下式极小化：

$$Q_N(\boldsymbol{\beta})=\begin{bmatrix}\frac{1}{N}\sum_i\mathbf{x}_iu_i\\ \frac{1}{N}\sum_i\mathbf{x}_iu_i^3\end{bmatrix}'\mathbf{W}_N\begin{bmatrix}\frac{1}{N}\sum_i\mathbf{x}_iu_i\\ \frac{1}{N}\sum_i\mathbf{x}_iu_i^3\end{bmatrix}\tag{6.2}$$

其中，$u_i=y_i-\mathbf{x}_i'\boldsymbol{\beta}$，$\mathbf{W}_N$ 表示 $2K\times 2K$ 阶加权矩阵。对于 $\mathbf{W}_N$ 的某些选择来说，这个估计量比 OLS 更有效。这种例子将在 6.3.6 节加以分析。

6.2.5 *工具变量回归*

工具变量估计是广义矩方法估计的一个重要例子。

考察线性回归模型 $y=\mathbf{x}'\boldsymbol{\beta}+u$，其复杂情况是，$\mathbf{x}$ 的某些元素与误差项相关，所以 OLS 关于 $\boldsymbol{\beta}$ 是非一致的。假定与 $\mathbf{x}$ 相关的工具(**instruments**)$\mathbf{z}$ 存在(已在 4.8 节介绍)，但要求满足 $\mathrm{E}[u|\mathbf{z}]=0$。那么，$\mathrm{E}[y-\mathbf{x}'\boldsymbol{\beta}|\mathbf{z}]=0$。利用类似于用于获得 OLS 例子(6.1)的代数运算，我们用 $\mathbf{z}$ 去乘，以便得到无条件的总体矩条件：

$$\mathrm{E}[\mathbf{z}(y-\mathbf{x}'\boldsymbol{\beta})]=\mathbf{0}\tag{6.3}$$

矩方法估计量就是求解相应的样本矩条件：

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i(y_i-\mathbf{x}_i'\boldsymbol{\beta})=\mathbf{0}$$

若 $\dim(\mathbf{z})=K$，则得到 $\hat{\boldsymbol{\beta}}_{\mathrm{MM}}=(\sum_i\mathbf{z}_i\mathbf{x}_i')^{-1}\sum_i\mathbf{z}_iy_i$，这是 4.8.6 节曾引进的线性工具变量估计量。

如果潜在工具比回归元个数多，使得 $\dim(\mathbf{z})>K$，并且方程个数多于未知数，就不存在解。一种可能性是使用恰好 K 个工具，但有效性却有损失。然而，广义矩方法估计量则是利用二次损失选择 $\hat{\boldsymbol{\beta}}$，以使向量 $N^{-1}\sum_i\mathbf{z}_i(y_i-\mathbf{x}_i'\boldsymbol{\beta})$ 尽可能小，所以 $\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}$ 对下式极小化：

$$Q_N(\boldsymbol{\beta}) = \left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})\right]'\mathbf{W}_N\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})\right] \tag{6.4}$$

其中，$\mathbf{W}_N$ 表示 $\dim(\mathbf{z})\times\dim(\mathbf{z})$的加权矩阵。2SLS 估计量(参见 4.8.6 节)对应于对 $\mathbf{W}_N$ 的特殊选择。

6.4 节对线性模型的工具变量方法进行了相当详细的阐述。广义矩方法的优点是，它提供设定加权矩阵 $\mathbf{W}_N$ 最优选择的一种方法，所得的估计量比 2SLS 更有效。

6.5 节涵盖非线性模型的工具变量方法。广义矩方法的一个优点是，对非线性回归的推广是简单易行的。于是，我们直接用非线性模型 $u=y-g(\mathbf{x}'\boldsymbol{\beta})$或 $u=r(y,\mathbf{x},\boldsymbol{\beta})$来代替前面关于 $Q_N(\boldsymbol{\beta})$的表达式中的 $y-\mathbf{x}'\boldsymbol{\beta}$。

6.2.6 面板数据

另一个广义矩方法的重要应用及相关的估计方法是面板数据回归。

举一个例子，假定 $y_{it}=\mathbf{x}_{it}'\boldsymbol{\beta}+u_{it}$，其中，$i$ 表示个体，t 表示时间。由 6.2.1 节知道，y_{it}对 $\mathbf{x}_{it}$的混合 OLS 回归，是建立在条件 $\mathrm{E}[\mathbf{x}_{it}u_{it}]=\mathbf{0}$ 基础之上的 MM 估计量。另外，假定误差 u_{it}与回归元在一些时期而非当前时期是不相关的。于是，对于 $s\neq t$，$\mathrm{E}[\mathbf{x}_{it}u_{it}]=\mathbf{0}$ 提供了能用于获得更有效估计量的另外矩条件。

第 22 章和第 23 章将提供用于面板数据内容的一些广义矩方法。

6.2.7 源于经济理论的矩条件

利用经济理论可以得到用于估计基础的矩条件。

这里以下述模型开始阐述：

$$y_t=\mathrm{E}[y_t\,|\,\mathbf{x}_t,\boldsymbol{\beta}]+u_t$$

其中，右边第一项测算出以 $\mathbf{x}$ 为条件的 y 的“预测”成分，第二项测算出“非预测”成分。举一个例子，y 可以表示资产收益或通货膨胀率。在理性预期和市场出清或市场有效这两个假设条件下，我们得到下述结果：非预测成分在确定 $\mathrm{E}[y\,|\,\mathbf{x}]$时，利用时间 t 获得的任何信息都是不可预测的。那么有：

$$\mathrm{E}[(y_t-\mathrm{E}[y_t\,|\,\mathbf{x}_t,\boldsymbol{\beta}])\,|\,\mathcal{I}_t]=0$$

其中，$\mathcal{I}_t$ 表示在时间 t 可利用的信息。

由期望迭代定律，$\mathrm{E}[\mathbf{z}_t(y_t-\mathrm{E}[y_t\,|\,\mathbf{x}_t,\boldsymbol{\beta}])]=0$，其中，$\mathbf{z}_t$ 表示由 $\mathcal{I}_t$ 的任何子集所形成的。由于信息集的任何部分都能用作工具，这就提供了可作为估计基础的许多矩条件。倘若使用时间序列数据，则广义矩方法最小化二次形式：

$$Q_T(\boldsymbol{\beta})=\left[\frac{1}{T}\sum\nolimits_{t=1}^{T}\mathbf{z}_tu_t\right]'\mathbf{W}_T\left[\frac{1}{T}\sum\nolimits_{t=1}^{T}\mathbf{z}_tu_t\right]$$

其中，$u_t=y_t-\mathrm{E}[y_t\,|\,\mathbf{x}_t,\boldsymbol{\beta}]$。倘若在单个时点 t 上使用横截面数据，则广义矩方法最小化二次形式：

$$Q_N(\boldsymbol{\beta}) = \left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i u_i\right]'\mathbf{W}_N\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i u_i\right]$$

其中，$u_i = y_i - \mathrm{E}[y_i|\mathbf{x}_i,\boldsymbol{\beta}]$，下标 t 因为仅有一个时期需要加以分析而被省略。

上述方法并没有受限于推导中所用的可加结构。所需要的全部内容是误差项 u_t 满足性质 $\mathrm{E}[u_t|\mathcal{I}_t]=0$。这类条件产生于欧拉条件，欧拉条件源于在确定性条件下的跨时期决策模型。例如，汉森和辛格尔顿(Hansen and Singleton, 1982)曾阐述的期望寿命效用最大化模型，导致了欧拉条件 $\mathrm{E}[u_t|\mathcal{I}_t]=0$，其中，$u_t=\beta g_{t+1}^{\alpha} r_{t+1}-1$，而 $g_{t+1}=c_{t+1}/c_t$ 表示两个时期的消费比，r_{t+1} 表示资产收益。参数 β 与 α 分别表示跨时期折现率与相对风险规避的参数，它们既可以是时间序列数据，又可以利用横截面数据通过广义矩方法得以估计，正如前面所做的那样，还有新定义的 u_t。汉森(Hansen, 1982)以及汉森和辛格尔顿(Hansen and Singleton, 1982)都考虑了时间序列数据，麦柯迪(MaCurdy, 1983)利用面板数据，对消费和劳动力供给进行了建模。

6.3 广义矩方法

本节将阐述广义矩方法估计的一般理论。广义矩方法定义出一类估计量。正如对分布的不同选取会产生各种不同的 ML 估计量一样，对矩条件与加权矩阵的不同选取也会产生各种不同的广义矩方法估计量。我们既讨论这些问题，又阐述估计广义矩方法估计量的方差矩阵方法，以及通常的一致性和渐近正态性质。

6.3.1 *矩方法估计量*

研究起点是假定存在 q 个参数的 r 个矩条件：

$$\mathrm{E}[\mathbf{h}(\mathbf{w}_i,\boldsymbol{\theta}_0)]=\mathbf{0} \tag{6.5}$$

其中，$\boldsymbol{\theta}$ 表示 $q\times1$ 维向量，$\mathbf{h}(\cdot)$ 表示 $r\times1$ 维向量函数，满足 $r\geq q$，而 $\boldsymbol{\theta}_0$ 表示数据生成过程中 $\boldsymbol{\theta}$ 的值。向量 $\mathbf{w}$ 包括所有可观测值，包含有关的因变量 $\mathbf{y}$、潜在内生回归元 $\mathbf{x}$ 以及工具变量 $\mathbf{z}$。因变量 $\mathbf{y}$ 可以是一个向量，因此，对方程组或面板数据的应用要进行归类。期望是关于 $\mathbf{w}$ 的所有随机成分的，由此也是关于 $\mathbf{y}$、$\mathbf{x}$ 和 $\mathbf{z}$ 的。

对 $\mathbf{h}(\cdot)$ 函数形式的选择，在性质上类似于对模型选择，而且会随着应用而变化。表 6.1 中的内容总结了 $\mathbf{h}(\mathbf{w})=\mathbf{h}(y,\mathbf{x},\mathbf{z},\boldsymbol{\theta})$ 的一些单方程例子，这已在 6.2 节中阐述过。

表 6.1　广义矩方法：例子

矩函数 $h(\cdot)$	估计方法
$y-\mu$	关于总体均值的矩方法
$\mathbf{x}(y-\mathbf{x}'\boldsymbol{\beta})$	普通最小二乘法回归
$\mathbf{z}(y-\mathbf{x}'\boldsymbol{\beta})$	工具变量回归
$\partial \ln f(y\|\mathbf{x},\boldsymbol{\theta})/\partial\boldsymbol{\theta}$	极大似然估计

只要 $r=q$,就能应用矩方法。总体矩等于零,可用相应样本矩等于零来代替,矩方法估计量 $\hat{\boldsymbol{\theta}}_{\mathrm{MM}}$(**method of moments estimator**)被定义为下式的解:

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{h}(\mathbf{w}_i,\hat{\boldsymbol{\theta}})=\mathbf{0} \tag{6.6}$$

这等价于对

$$Q_N(\boldsymbol{\theta})=\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{h}(\mathbf{w}_i,\boldsymbol{\theta})\right]'\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{h}(\mathbf{w}_i,\boldsymbol{\theta})\right]$$

求极小值的估计方程估计量,其渐近分布已由 5.4 节阐述,并由 6.3.3 节的式(6.13)重新描述。

6.3.2 *广义矩方法估计量*

广义矩方法估计量建立在 r 个独立的矩条件(6.5)的基础上,并且有 q 个参数被估计。

如果 $r=q$,那么模型称为恰好识别的(**just-identified**),并可使用矩方法估计量(6.6)。更正式地讲,$r=q$ 仅仅是恰好识别的必要条件,而且我们另外要求命题 6.1 中的 $\mathbf{G}_0$ 具有秩 q。识别将在 6.3.9 节中加以讨论。

如果 $r>q$,那么模型称为过度识别的(**overidentified**),由于方程个数(r)比未知数个数(q)多,所以对于 $\hat{\boldsymbol{\theta}}$ 来说,式(6.6)没有解。相反,选取 $\hat{\boldsymbol{\theta}}$ 以使二次形式 $N^{-1}\sum_i\mathbf{h}(\mathbf{w}_i,\hat{\boldsymbol{\theta}})$尽可能地接近于零。特别地,广义矩方法估计量(**generalized methods of moments estimator**)$\hat{\boldsymbol{\theta}}_{\mathrm{GMM}}$就是对目标函数

$$Q_N(\boldsymbol{\theta})=\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{h}(\mathbf{w}_i,\boldsymbol{\theta})\right]'\mathbf{W}_N\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{h}(\mathbf{w}_i,\boldsymbol{\theta})\right] \tag{6.7}$$

求极小值,其中,$r\times r$ 阶加权矩阵 $\mathbf{W}_N$ 表示对称正定的、可能具有有限概率极限的随机形式,但不依赖 $\boldsymbol{\theta}$。$\mathbf{W}_N$ 的下标 N 表示,它的值可能依赖于样本。然而,当 $N\to\infty$时,$\mathbf{W}_N$ 的维数 r 是固定的。目标函数还能用矩阵记号表示成 $Q_N(\boldsymbol{\theta})=N^{-1}\mathbf{l}'\mathbf{H}(\boldsymbol{\theta})\times\mathbf{W}_N\times N^{-1}\mathbf{H}(\boldsymbol{\theta})'\mathbf{l}$,其中,$\mathbf{l}$ 表示 $N\times 1$ 维向量,而 $\mathbf{H}(\boldsymbol{\theta})$表示 $N\times r$ 阶矩阵,其第 i 行是 $\mathbf{h}(y_i,\mathbf{x}_i,\boldsymbol{\theta})'$。

对加权矩阵 $\mathbf{W}_N$ 的不同选择,将会产生各种不同估计量,如果 $r>q$,那么估计量是一致的,但具有不同的方差。一种简单选择是设 $\mathbf{W}_N$ 为单位矩阵,尽管这常常是差的选择。于是,$Q_N(\boldsymbol{\theta})=\bar{h}_1^2+\bar{h}_2^2+\cdots+\bar{h}_r^2$ 表示 r 个样本平均平方之和,其中,$\bar{h}_j=N^{-1}\sum_i h_j(\mathbf{w}_i,\boldsymbol{\theta})$,而 $h_j(\cdot)$表示 $\mathbf{h}(\cdot)$的第 j 个成分。6.3.5 节给出对 $\mathbf{W}_N$ 的最优选择。

求式(6.7)中 $Q_N(\boldsymbol{\theta})$关于 $\boldsymbol{\theta}$ 的微分,得到广义矩方法一阶条件:

$$\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\partial\mathbf{h}_i(\hat{\boldsymbol{\theta}})'}{\partial\boldsymbol{\theta}}\bigg|_{\hat{\theta}}\right]\times\mathbf{W}_N\times\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{h}_i(\hat{\boldsymbol{\theta}})\right]=\mathbf{0} \tag{6.8}$$

其中,$\mathbf{h}_i(\boldsymbol{\theta})=\mathbf{h}_i(\mathbf{w}_i,\boldsymbol{\theta})$,并且我们用尺度因子[1](scaling factor)1/2 去乘。一般来

[1] 又称为标度因子。——译者注

说，这些方程十分复杂且关于 $\hat{\boldsymbol{\theta}}$ 是非线性的，因为 $\hat{\boldsymbol{\theta}}$ 既出现在第一项中也出现在第三项中。数值求解法将在第 10 章加以阐述。

6.3.3 *广义矩方法估计量分布*

广义矩方法估计量的渐近分布是以下述命题形式给出的，对它的推导则由 6.3.9 节给出。

命题 6.16(广义矩方法估计量分布) 做出下述假设：

(i) 对矩条件(6.5)施加数据生成过程；即 $\mathrm{E}[\mathbf{h}(\mathbf{w},\boldsymbol{\theta}_0)=\mathbf{0}$。

(ii) $r\times 1$ 维向量函数 $\mathbf{h}(\cdot)$ 满足 $\mathbf{h}(\mathbf{w},\boldsymbol{\theta}^{(1)})=\mathbf{h}(\mathbf{w},\boldsymbol{\theta}^{(2)})$，当且仅当 $\boldsymbol{\theta}^{(1)}=\boldsymbol{\theta}^{(2)}$。

(iii) 下面的 $r\times q$ 阶矩阵存在且是有限的，其秩为 q：

$$\mathbf{G}_0=\operatorname{plim}\frac{1}{N}\sum_{i=1}^{N}\left[\left.\frac{\partial\mathbf{h}_i}{\partial\boldsymbol{\theta}'}\right|_{\theta_0}\right]\tag{6.9}$$

(iv) $\mathbf{W}_N\xrightarrow{P}\mathbf{W}_0$，其中，$\mathbf{W}_0$ 表示有限对称正定矩阵。

(v) $N^{-1/2}\sum_{i=1}^{N}\mathbf{h}_i|_{\theta_0}\xrightarrow{d}\mathcal{N}[\mathbf{0},\ \mathbf{S}(\boldsymbol{\theta}_0)]$，其中：

$$\mathbf{S}_0=\operatorname{plim}N^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N}[\mathbf{h}_i\mathbf{h}_j'|_{\theta_0}]\tag{6.10}$$

那么，广义矩方法估计量 $\hat{\boldsymbol{\theta}}_{\mathrm{GMM}}$被定义为由式(6.8)给出的一阶条件的根，此估计量关于 $\boldsymbol{\theta}_0$ 是一致的，并且：

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{\mathrm{GMM}}-\boldsymbol{\theta}_0)\xrightarrow{d}\mathcal{N}[\mathbf{0},(\mathbf{G}_0'\mathbf{W}_0\mathbf{G}_0)^{-1}(\mathbf{G}_0'\mathbf{W}_0\mathbf{S}_0\mathbf{W}_0\mathbf{G}_0)(\mathbf{G}_0'\mathbf{W}_0\mathbf{G}_0)^{-1}]\tag{6.11}$$

一些重要的专门研究是下述内容。

首先，在微观经济计量分析中，通常假定数据对于不同 i 是独立的，所以式(6.10)简化为：

$$\mathbf{S}_0=\operatorname{plim}\frac{1}{N}\sum_{i=1}^{N}[\mathbf{h}_i\mathbf{h}_i'|_{\theta_0}]\tag{6.12}$$

另外，若假定数据是同分布的，则式(6.9)与式(6.10)简化为 $\mathbf{G}_0=\mathrm{E}[\partial\mathbf{h}/\partial\boldsymbol{\theta}'|_{\theta_0}]$与 $\mathbf{S}_0=\mathrm{E}[\mathbf{h}\mathbf{h}'|_{\theta_0}]$，这种记号已被许多作者使用。

其次，在 $r=q$ 的恰好识别情况下，对于包含 ML 与 LS 的许多估计量的情况来说，结果可简化成由 5.4 节阐述的估计方程估计量的那些结果。为了理解这一点，注意到，当 $r=q$ 时，矩阵 $\mathbf{G}_0$、$\mathbf{W}_0$ 和 $\mathbf{S}_0$ 都是可逆方阵，所以$(\mathbf{G}_0'\mathbf{W}_0\mathbf{G}_0)^{-1}=\mathbf{G}_0^{-1}\mathbf{W}_0^{-1}(\mathbf{G}_0')^{-1}$，并且式(6.11)中的方差矩阵可以简化。由此可得，对于式(6.6)中的 MM 估计量来说，有：

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{\mathrm{MM}}-\boldsymbol{\theta}_0)\xrightarrow{d}\mathcal{N}[\mathbf{0},\mathbf{G}_0^{-1}\mathbf{S}_0(\mathbf{G}_0')^{-1}]\tag{6.13}$$

像广义矩方法估计量一样，GMM 估计量总是可以被计算出，而且对满秩加权矩阵选择来说是不变的。

再次，对矩阵 $\mathbf{W}_N$ 的最佳选择是使得 $\mathbf{W}_0=\mathbf{S}_0^{-1}$。那么，在式(6.11)中的方差矩阵可简化为$(\mathbf{G}_0'\mathbf{S}_0^{-1}\mathbf{G}_0)^{-1}$。这将在 6.3.5 节中详细阐述。

6.3.4 *方差矩阵估计*

有关广义矩方法估计量的统计推断可能是式(6.11)中的 $\mathbf{G}_0$、$\mathbf{W}_0$ 和 $\mathbf{S}_0$ 的一致估计值 $\hat{\mathbf{G}}$、$\hat{\mathbf{W}}$、$\hat{\mathbf{S}}$。在相对弱分布的假设下，很容易获得一致估计值。

对 $\mathbf{G}_0$ 来说，一个明显估计量是：

$$\hat{\mathbf{G}} = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \mathbf{h}_i}{\partial \boldsymbol{\theta}'}\Big|_{\hat{\theta}} \tag{6.14}$$

对 $\mathbf{W}_0$ 而言，可使用简单加权矩阵。关于 $r\times r$ 阶矩阵 $\mathbf{S}_0$ 的估计量，会随着做出有关数据生成过程的随机假设而变化。通常，微观经济计量分析假定，对于不同 i 具有独立性，因而 $\mathbf{S}_0$ 具有比较简单的形式(6.10)。于是，一个明显估计量是：

$$\hat{\mathbf{S}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{h}_i(\hat{\boldsymbol{\theta}})\mathbf{h}_i(\hat{\boldsymbol{\theta}})' \tag{6.15}$$

由于 $\mathbf{h}(\cdot)$表示 $r\times 1$ 维的，所以 $\mathbf{S}_0$ 中至多存在独一无二的 $r(r+1)/2$ 个需要估计的有限数。因此，假定 $\mathrm{E}[\mathbf{h}_i,\mathbf{h}_i']$存在并依赖于少数几个参数，而不需要对方差 $\mathrm{E}[\mathbf{h}_i\mathbf{h}']$参数化，当 $N\to\infty$时，$\hat{\mathbf{S}}$ 是一致的。所需要的全部内容就是，添加某种合适的附加假设，以确保 $\mathrm{plim}\, N^{-1}\sum_i \hat{\mathbf{h}}_i\hat{\mathbf{h}}_i' = \mathrm{plim}\, N^{-1}\sum_i \mathbf{h}_i\mathbf{h}_i'$。例如，如果 $\hat{\mathbf{h}}_i = \mathbf{x}_i\hat{u}_i$，其中，$\hat{u}_i$ 表示 OLS 残差，我们由 4.4 节知道，需要假定该估计量的四阶矩存在。

对这些结果加以综合，就得出广义矩方法估计量服从渐近正态分布，其均值为 $\boldsymbol{\theta}_0$，而估计渐近方差为：

$$\hat{\mathrm{V}}[\hat{\boldsymbol{\theta}}_{\mathrm{GMM}}] = \frac{1}{N}(\hat{\mathbf{G}}'\mathbf{W}_N\hat{\mathbf{G}})^{-1}\hat{\mathbf{G}}'\mathbf{W}_N\hat{\mathbf{S}}\mathbf{W}_N\hat{\mathbf{G}}(\hat{\mathbf{G}}'\mathbf{W}_N\hat{\mathbf{G}})^{-1} \tag{6.16}$$

这个方差矩阵估计量是稳健的估计量，即艾克—怀特异方差一致估计量关于最小二乘法估计量的扩展。

人们还可取期望，同时对 $\mathbf{G}_0$ 使用 $\hat{\mathbf{G}}_{\mathrm{E}} = N^{-1}\sum_i \mathrm{E}[\partial\mathbf{h}_i/\partial\boldsymbol{\theta}']|_{\hat{\theta}}$，而对 $\mathbf{S}_0$ 使用 $\hat{\mathbf{S}}_{\mathrm{E}} = N^{-1}\sum_i \mathrm{E}[\mathbf{h}_i\mathbf{h}_i']|_{\hat{\theta}}$。不过，为了取期望，通常需要附加的分布假设，而方差矩阵估计关于分布错误设定将不是稳健的。

在时间序列情况下，$\mathbf{h}_t$ 以时间 t 表示下标，而渐近理论建立在时期 $T\to\infty$的基础上。就时间数据而言，$\mathbf{h}_t$ 是一个向量 MA(q)过程，$\mathrm{V}[\hat{\boldsymbol{\theta}}_{\mathrm{GMM}}]$的通常估计量由纽韦和韦斯特(Newey and West, 1987b)提出，他们使用式(6.16)与 $\hat{\mathbf{S}} = \hat{\boldsymbol{\Omega}}_0 + \sum_{j=1}^{q}\left(1-\frac{j}{q+1}\right)(\hat{\boldsymbol{\Omega}}_j + \hat{\boldsymbol{\Omega}}_j')$，其中，$\hat{\boldsymbol{\Omega}} = T^{-1}\sum_{t=j+1}^{T}\hat{\mathbf{h}}_t\hat{\mathbf{h}}_{t-j}'$。除了同期相关之外，还允许 $\mathbf{h}_t$ 中时间序列相关。关于协方差矩阵估计的进一步详细内容，包括时间序列情况下的一些改进，已由戴维森和麦金农(Davidson and MacKinnon, 1993, 17.5 节)、哈密尔顿(Hamilton, 1994)以及哈恩和莱文(Haan and Levin, 1997)给出。

6.3.5 *最优加权矩阵*

运用广义矩方法，需要对式(6.7)的矩函数 $\mathbf{h}(\cdot)$和加权矩阵 $\mathbf{W}_N$ 进行设定。

容易选取 $\mathbf{W}_N$，以便获得给定设定函数 $\mathbf{h}(\cdot)$时，具有最小渐近方差的广义矩方

法估计量。这常常称为最优广义矩方法，尽管它是最优性的一个受限形式，因为如果对 $\mathbf{h}(\cdot)$ 选择不好，则会产生非常无效的估计量。

对恰好识别模型来说，就任何满秩加权矩阵而言，可获得同样的估计量（矩方法估计量），因此，人们还是最好令 $\mathbf{W}_N=\mathbf{I}_q$。

对满足 $r>q$ 的过度识别模型且 $\mathbf{S}_0$ 为已知的情况来说，通过选择加权矩阵 $\mathbf{W}_N=\mathbf{S}_0^{-1}$ 来获得最有效的广义矩方法估计量。于是，对命题中给出的方差矩阵可进行简化，并且：

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{\mathrm{GMM}}-\boldsymbol{\theta}_0)\xrightarrow{d}\mathcal{N}[\mathbf{0},(\mathbf{G}_0'\mathbf{S}_0^{-1}\mathbf{G}_0)^{-1}] \tag{6.17}$$

此结果归功于汉森（Hansen，1982）。

这一结果可利用类似于线性模型中建立 GLS 是最有效的 WLS 估计量的那些矩阵推理来获得。甚至更简单的是，人们能直接对目标函数进行推导。对二次形式 $\mathbf{u}'\mathbf{W}\mathbf{u}$ 求极小值的 LS 估计量来说，最有效的估计量是设 $\mathbf{W}=\boldsymbol{\Sigma}^{-1}=\mathrm{V}[\mathbf{u}]^{-1}$ 的 GLS。在式(6.7)中，广义矩方法目标函数是满足 $\mathbf{u}=N^{-1}\sum_i\mathbf{h}_i(\boldsymbol{\theta})$ 的二次形式，所以最优 $\mathbf{W}=(\mathrm{V}[N^{-1}\sum_i\mathbf{h}_i(\boldsymbol{\theta})])^{-1}=\mathbf{S}_0^{-1}$。最优广义矩方法估计量可通过样本矩条件下方差矩阵的逆来进行加权。

最优 GMM

在实际应用中，$\mathbf{S}_0$ 是未知的，并且我们设 $\mathbf{W}_N=\hat{\mathbf{S}}^{-1}$，其中，$\hat{\mathbf{S}}$ 表示关于 $\mathbf{S}_0$ 是一致的。最优广义矩方法估计量能利用两步法[1]（**two-step procedure**）来获得。第一步，广义矩方法估计量可利用对 $\mathbf{W}_N$ 的次优选择，比如为了简单起见，取 $\mathbf{W}_N=\mathbf{I}_r$。第一步，利用式(6.15)估计 $\hat{\mathbf{S}}$。第二步，利用最优加权矩阵 $\mathbf{W}_N=\hat{\mathbf{S}}^{-1}$，得到一个最优 GMM 估计量。

于是，最优广义矩方法估计量（**optimal GMM estimator**）或两步广义矩方法估计量（**two-step GMM estimator**）$\hat{\boldsymbol{\theta}}_{\mathrm{OGMM}}$，建立在 $\mathbf{h}_i(\boldsymbol{\theta})$ 对

$$Q_N(\boldsymbol{\theta})=\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{h}_i(\boldsymbol{\theta})\right]'\hat{\mathbf{S}}^{-1}\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{h}_i(\boldsymbol{\theta})\right] \tag{6.18}$$

求极小值的基础上。其极限分布已由式(6.17)给出。最优广义矩方法估计量服从渐近正态分布，均值为 $\boldsymbol{\theta}_0$，而估计渐近方差有相对简单的公式：

$$\mathrm{V}[\hat{\boldsymbol{\theta}}_{\mathrm{OGMM}}]=N^{-1}(\hat{\mathbf{G}}'\tilde{\mathbf{S}}^{-1}\hat{\mathbf{G}})^{-1} \tag{6.19}$$

通常，$\hat{\mathbf{G}}$ 与 $\tilde{\mathbf{S}}$ 的计算均在 $\hat{\boldsymbol{\theta}}_{\mathrm{OGMM}}$ 处进行，所以 $\tilde{\mathbf{S}}$ 使用与 $\hat{\mathbf{S}}$ 相同的公式，只是在 $\hat{\boldsymbol{\theta}}_{\mathrm{OGMM}}$ 处进行计算。一种可供选择的方法，是在第一步估计量处加以计算，如同使用 $\boldsymbol{\theta}_0$ 的任何一致估计量一样。

值得注意的是，为了估计次优广义矩方法的方差矩阵，式(6.18)中的最优广义矩方法估计量并不要求附加超出其需要，允许使用式(6.16)的那些随机假设。在这两种情况下，要求 $\hat{\mathbf{S}}$ 关于 $\mathbf{S}_0$ 是一致的，并且由式(6.15)之后的讨论可知，这需要几个附加假设。当误差是异方差时，这完全可以与为使 GLS 比 OLS 更有效而需

[1] 这里，把 procedure 译成方法，它还有程序之意。——译者注

要的附加假设形成对比。然而，误差中的异方差性将会影响到对 $\mathbf{h}_i(\boldsymbol{\theta})$ 的最优选择（参见 6.3.7 节）。

两步广义矩方法小样本偏倚

对过度识别模型来说，理论研究表明，最好的方法是使用最优广义矩方法。不过，在具体实施时，从理论上来说，最优加权矩阵 $\mathbf{W}_N=\mathbf{S}_0^{-1}$ 需要用一致估计值 $\hat{\mathbf{S}}^{-1}$ 来代替。这种代替在渐近形式上不会造成什么差异，但它在有限样本上产生差异。尤其是，使式(6.18)中的 $\mathbf{h}_i(\boldsymbol{\theta})$ 增大的个体观测值，可能会增大式(6.18)中的 $\hat{\mathbf{S}}=N^{-1}\sum_i\hat{\mathbf{h}}_i\hat{\mathbf{h}}_i'$，导致 $N^{-1}\sum_i\mathbf{h}_i(\boldsymbol{\theta})$ 与 $\hat{\mathbf{S}}$ 相关。注意到，由于取概率极限，所以类似地，$\mathbf{S}_0=\operatorname{plim} N^{-1}\sum_i\mathbf{h}_i\mathbf{h}_i'$ 并没有受到影响。

奥尔顿金和西格尔(Altonji and Segal, 1996)在运用面板数据对协方差结构模型进行估计时，已经阐述过这个问题(参见 22.5 节)。他们使用有关的最小距离估计量(参见 6.7 节)，但是文献却把他们的结果解释成与利用横截面数据或短面板的广义矩方法估计有关。如人们所料，在模拟研究中，最优估计量比一步估计量更有效。不过，最优估计量具有的有限样本偏倚如此之大，以致它的均方误差根远大于一步估计量的均方误差根。

奥尔顿金和西格尔(Altonji and Segal, 1996)还提供了一种变形，即独立加权最优(**independently weighted optimal**)估计量，它是利用观测值而不是样本矩构造加权矩阵。他们把样本分成 G 个组，一种明显选择是 $G=2$，并且对下式极小化：

$$Q_N(\boldsymbol{\theta})=\frac{1}{G}\sum_g\mathbf{h}_g(\boldsymbol{\theta})\hat{\mathbf{S}}_{(-g)}^{-1}\mathbf{h}_g(\boldsymbol{\theta}) \tag{6.20}$$

其中，$\mathbf{h}_g(\boldsymbol{\theta})$ 表示对第 g 个组计算，而 $\hat{\mathbf{S}}_{(-g)}$ 表示利用除了第 g 个组之外的所有组计算。这个估计量偏倚很小，因为由构造知，加权矩阵 $\hat{\mathbf{S}}_{(-g)}^{-1}$ 与 $\mathbf{h}_g(\boldsymbol{\theta})$ 是独立的。然而，分割样本会导致有效性损失。相反，霍罗威茨(Horowitz, 1998a)使用自助法(参见 11.6.4 节)

在奥尔顿金和西格尔(Altonji and Segal, 1996)的例子中，$\mathbf{h}_i$ 涉及二阶矩，所以 $\hat{\mathbf{S}}$ 涉及四阶矩。在其他例子中，最优估计量的有限样本问题不是显著的，其中，$\mathbf{h}_i$ 仅仅包含一阶矩。不过，奥尔顿金和西格尔的结果建议，在利用最优广义矩方法时要小心谨慎，而且一步广义矩方法与最优广义矩方法估计值之间的差异或许表明，最优广义矩方法存在有限样偏倚的问题。

矩约束的个数

通常，进一步增加矩约束会改进渐近有效性，这样做减少了最优广义矩方法估计量的极限方差$(\mathbf{G}_0'\mathbf{S}_0^{-1}\mathbf{G}_0)^{-1}$，或者最不利的情况就是，渐近有效性尚未改变。

进一步增加矩条件的益处会随着应用而变化。例如，若估计量是 MLE，则 MLE 是完全有效的，所以并未得到任何改进。因为作为工具的变量可能与许多工具的组合比其与单个工具更为高度相关，所以文献专注于那种值得考虑的工具变量估计。

然而，由于矩约束的个数不能大于观测值的个数，所以存在限制性。另外，增加更多矩条件，会增大有限样本偏倚的似然，相关问题类似于线性模型弱工具的那些问题(参见 4.9 节)。斯托克等人(Stock et al., 2002)简要考虑了非线性模型中的弱工具。

6.3.6 带有对称误差回归的例子

为了阐明广义矩方法渐近结果，我们回到6.2.4节引进的附加矩约束例子上。对这个例子来说，β_{GMM}的目标函数已由式(6.2)给出。所需要做的全部内容就是对 $\mathbf{W}_N$ 进行设定，例如，$\mathbf{W}_N=\mathbf{I}$。

为了获得这个估计量的分布，我们使用6.3节的一般记号。把式(6.5)中的函数 $\mathbf{h}(\cdot)$ 专门化为：

$$\mathbf{h}(\mathbf{y},\mathbf{x},\boldsymbol{\beta})=\begin{bmatrix}\mathbf{x}(y-\mathbf{x}'\boldsymbol{\beta})\\ \mathbf{x}(y-\mathbf{x}'\boldsymbol{\beta})^3\end{bmatrix}\Rightarrow\frac{\partial\mathbf{h}(\mathbf{y},\mathbf{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}'}=\begin{bmatrix}-\mathbf{x}\mathbf{x}'\\ -3\mathbf{x}\mathbf{x}'(y-\mathbf{x}'\boldsymbol{\beta})^2\end{bmatrix}$$

这些表达式利用式(6.9)与式(6.12)直接推导出关于 $\mathbf{G}_0$ 与 $\mathbf{S}_0$ 的表达式，因此，由式(6.14)与式(6.15)得出一致估计值：

$$\hat{\mathbf{G}}=\begin{bmatrix}-\frac{1}{N}\sum_i\mathbf{x}_i\mathbf{x}_i'\\ -\frac{1}{N}\sum_i3\hat{u}_i^2\mathbf{x}_i\mathbf{x}_i'\end{bmatrix}\tag{6.21}$$

与

$$\hat{\mathbf{S}}=\begin{bmatrix}\frac{1}{N}\sum_i\hat{u}_i^2\mathbf{x}_i\mathbf{x}_i' & \frac{1}{N}\sum_i\hat{u}_i^4\mathbf{x}_i\mathbf{x}_i'\\ \frac{1}{N}\sum_i\hat{u}_i^4\mathbf{x}_i\mathbf{x}_i' & \frac{1}{N}\sum_i\hat{u}_i^6\mathbf{x}_i\mathbf{x}_i'\end{bmatrix}\tag{6.22}$$

其中，$\hat{u}_i=y_i-\mathbf{x}_i'\hat{\boldsymbol{\beta}}$。可供选择的估计值通过先在 $\mathbf{G}_0$ 与 $\mathbf{S}_0$ 处计算期望来获得，但这要求 $\mathrm{E}[u^2\mid\mathbf{x}]$、$\mathrm{E}[u^4\mid\mathbf{x}]$ 以及 $\mathrm{E}[u^6\mid\mathbf{x}]$ 存在的假设。把 $\hat{\mathbf{G}}$、$\hat{\mathbf{S}}$ 以及 $\mathbf{W}_N$ 代入式(6.16)，得到 $\hat{\boldsymbol{\beta}}_{GMM}$的估计渐近方差矩阵。

现在，考察带有最优加权矩阵的广义矩方法。这又一次对式(6.2)极小化，但是，从式(6.18)开始，有 $\mathbf{W}_N=\hat{\mathbf{S}}^{-1}$，其中，$\hat{\mathbf{S}}$ 已由式(6.22)定义。对 $\hat{\mathbf{S}}$ 进行计算，需要 $\hat{\boldsymbol{\beta}}$ 的一步一致估计值。一种明显的选择是满足 $\mathbf{W}_N=\mathbf{I}$ 的广义矩方法。在此例子中，OLS估计量是一致的，而且还可以被使用。利用式(6.19)，得到这种两步估计量的估计渐近方差矩阵 $\hat{\mathrm{V}}[\hat{\boldsymbol{\beta}}_{OGMM}]$，它等于：

$$\left(\begin{bmatrix}\sum_i\tilde{u}_i\mathbf{x}_i\mathbf{x}_i'\\ \sum_i\tilde{u}_i^3\mathbf{x}_i\mathbf{x}_i'\end{bmatrix}'\begin{bmatrix}\sum_i\tilde{u}_i^2\mathbf{x}_i\mathbf{x}_i' & \sum_i\tilde{u}_i^4\mathbf{x}_i\mathbf{x}_i'\\ \sum_i\tilde{u}_i^4\mathbf{x}_i\mathbf{x}_i' & \sum_i\tilde{u}_i^6\mathbf{x}_i\mathbf{x}_i'\end{bmatrix}^{-1}\begin{bmatrix}\sum_i\tilde{u}_i\mathbf{x}_i\mathbf{x}_i'\\ \sum_i\tilde{u}_i^3\mathbf{x}_i\mathbf{x}_i'\end{bmatrix}\right)^{-1}$$

其中，$\tilde{u}_i=y_i-\mathbf{x}_i'\hat{\boldsymbol{\beta}}_{OGMM}$，而各项用 N 去除就可以抵消。

在本例中，最优广义矩方法提高有效性的解析结果，可通过对非回归元的专门化而很容易地获得，其中，y 表示iid的，其均值为 μ。进一步地，假定 y 表示拉普拉斯分布，其标度参数等于1，在此情况下，密度是 $f(y)=(1/2)\times\exp\{-|y-\mu|\}$，$\mathrm{E}[y]=\mu$，$\mathrm{V}[y]=2$。同时，对奇数 r 来说，较高阶的中心矩 $\mathrm{E}[(y-\mu)^r]$等于0；而对偶数 r 来说，它等于 $r!$。样本中位数是完全有效的，这是因为它是MLE，并且可以证明，它具有渐近方差 $1/N$。样本均值是 $\bar{y}$ 无效的，其方差 $\mathrm{V}[\bar{y}]=\mathrm{V}[y]/N=2/N$。建立在两个矩条件 $\mathrm{E}[(y-\mu)]=0$ 与 $\mathrm{E}[(y-\mu)^3]=0$ 基础上的最优广义矩

方法估计量$\hat{\mu}_{opt}$具有下述的加权矩阵，即把更小权数放在二阶矩条件上，因为它具有相对大的方差，并且具有负的非对角线元素。可以证明，最优广义矩方法估计量$\hat{\mu}_{OGMM}$具有渐近方差 1.714 3/N(参见习题 6.3)。因此，尽管它与样本中位数相比，其有效性还很差，但是它比样本均值(方差为 2/N)更有效。

对这个例子来说，单位矩阵是加权矩阵的一个例外的差选择。它把太大的权数放在二阶矩条件上，得到一个 μ 的次优广义矩方法估计量，其渐近方差 19.14/N 比平稳的 $V[\bar{y}]=2/N$ 大出了许多倍。详细内容参见习题 6.3。

6.3.7 最优矩条件

6.3.5 节已给出令人惊奇的结果。从本质上讲，最优广义矩方法的要求并不比没有最优加权矩阵的广义矩方法所需要的假设多。不过，这种最优非常有局限性，因为它是以对式(6.15)或式(6.18)中矩函数 $\mathbf{h}(\cdot)$的选取为条件的。

广义矩方法定义出一类估计量，对 $\mathbf{h}(\cdot)$的各种不同选择对应于此类不同情况。对 $\mathbf{h}(\cdot)$的某些选择比另一些选择要好一些，这依赖于附加随机假设。例如，当误差是异方差时，$\mathbf{h}_i=\mathbf{x}_i u_i$ 会产生 OLS 估计量，而 $\mathbf{h}_i=\mathbf{x}_i u_i/V[u_i\,|\,\mathbf{x}_i]$则产生 GLS 估计量。这种选择 $\mathbf{h}(\cdot)$的潜在多样性尤其导致任何特殊的广义矩方法估计量。然而，在 m 估计中，就选择而言，从性质上来看，必须做出类似决策，例如，对误差平方和极小化，而不是对误差加权平方和或误差绝对离差和极小化。

如果做出完全分布假设，那么最有效的估计量是 MLE。因此，对式(6.5)中 $\mathbf{h}(\cdot)$的最优选择是：

$$\mathbf{h}(\mathbf{w},\boldsymbol{\theta})=\frac{\partial \ln f(\mathbf{w},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

其中，$f(w,\boldsymbol{\theta})$表示 $\mathbf{w}$ 的联合密度。对于具有因变量 $\mathbf{y}$ 与回归元 $\mathbf{x}$ 的回归来说，这是一个基于 $\mathbf{y}$ 与 $\mathbf{x}$ 的无条件联合密度 $f(\mathbf{y},\mathbf{x},\boldsymbol{\theta})$的无条件 MLE。在许多应用中，$f(\mathbf{y},\mathbf{x},\boldsymbol{\theta})=f(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})g(\mathbf{x})$，其中，$\mathbf{x}$ 的边际密度(未用的)参数不依赖于关注 $\boldsymbol{\theta}$ 的参数。于是，这正像使用基于条件密度 $f(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})$的条件 MLE 一样有效。这能用于矩方法估计或者带有加权矩阵 $\mathbf{W}_N=\mathbf{I}_q$ 的广义矩方法估计的基础，尽管任何满秩矩阵 $\mathbf{W}_N$ 也将给出 MLE。然而，由于广义矩方法估计的目的是为了避免做出全部分布假设的集合，所以这种结果在实际应用中很有限。

当做出不完全分布假设时，一个普通的起点是对条件的矩条件(**conditional moment condition**)进行设定。对于模型误差，诸如 $E[u|\mathbf{x}]=0$ 或 $E[u|\mathbf{z}]=0$ 来说，这通常是低阶矩条件。此条件的矩条件能导致可作为广义矩方法估计基础的许多无条件矩条件(**unconditional moment conditions**)，譬如 $E[\mathbf{z}u]=\mathbf{0}$。纽韦(Newey, 1990a, 1993)曾获得对于在不同 i 上为独立的数据的无条件的矩条件最优选择结果。

特别地，以 s 个条件的矩条件约束开始，有：

$$E[\mathbf{r}(y,\mathbf{x},\boldsymbol{\theta}_0)\,|\,\mathbf{z}]=\mathbf{0} \tag{6.23}$$

其中，$\mathbf{r}(\cdot)$表示 6.2.2 节引进的残差型 $s\times 1$ 维向量函数。一个纯量例子是 $E[y-\mathbf{x}'\boldsymbol{\theta}_0\,|\,\mathbf{z}]=0$。这里使用了工具变量记号，其中，$\mathbf{x}$ 表示回归元，某些 $\mathbf{x}$ 是潜在内生的，$\mathbf{z}$ 表示包括 $\mathbf{x}$ 的外生元素的工具。在没有内生性的比较简单模型中，有 $\mathbf{z}=\mathbf{x}$。

建立在式(6.2.3)基础上的 q 个参数 $\boldsymbol{\theta}$ 的广义矩方法估计是不可行的，因为典型地讲，仅仅存在几个条件的矩约束，而且情况经常如此，所以 $s \leqslant q$。相反，我们引进一个 $r \times s$ 阶矩阵函数工具 $\mathbf{D}(\mathbf{z})$，其中，$r \geqslant q$，同时注意到，期望迭代定律 $\mathrm{E}[\mathbf{D}(\mathbf{z})\mathbf{r}(y,\mathbf{x},\boldsymbol{\theta}_0)]=\mathbf{0}$ 能用作广义矩方法估计的基础。可以证明，矩阵函数 $\mathbf{D}(\mathbf{z})$ 的最优工具(**optimal instruments**)或最优选择是 $q \times s$ 阶矩阵：

$$\mathbf{D}^*(\mathbf{z},\boldsymbol{\theta}_0)=\mathrm{E}\left[\frac{\partial \mathbf{r}(y,\mathbf{x},\boldsymbol{\theta}_0)'}{\partial \boldsymbol{\theta}}\bigg|\mathbf{z}\right]\{\mathrm{V}[\mathbf{r}(y,\mathbf{x},\boldsymbol{\theta}_0)|\mathbf{z}]\}^{-1} \tag{6.24}$$

例如，戴维森和麦金农(Davidson and MacKinnon, 1993，第 604 页)就曾经给出一种推导。该最优工具矩阵 $\mathbf{D}^*(\mathbf{z})$ 表示 $q \times s$ 阶矩阵，因此，无条件的矩条件 $\mathrm{E}[\mathbf{D}^*(\mathbf{z})\ \mathbf{r}(y,\mathbf{x},\boldsymbol{\theta}_0)]=\mathbf{0}$ 刚好产生与参数同样多的矩条件。最优广义矩方法估计量直接求解与之相对应的样本矩条件：

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{D}^*(\mathbf{z}_i,\boldsymbol{\theta})\mathbf{r}(y_i,\mathbf{x}_i,\boldsymbol{\theta})=\mathbf{0} \tag{6.25}$$

最优估计量需要额外的假设，即用于形成式(6.24)中的 $\mathbf{D}^*(\mathbf{z},\boldsymbol{\theta}_0)$ 期望，而且具体落实要求用已知参数代替未知参数，所以使用了生成回归元。

例如，如果 $r(y,\mathbf{x},\boldsymbol{\theta})=y-\exp(\mathbf{x}'\boldsymbol{\theta})$，那么 $\partial r/\partial\boldsymbol{\theta}=\exp(\mathbf{x}'\boldsymbol{\theta})\mathbf{x}$，而式(6.24)要求对 $\mathrm{E}[\exp(\mathbf{x}'\boldsymbol{\theta}_0)\mathbf{x}|\mathbf{z}]$ 与 $\mathrm{V}[y-\exp(\mathbf{x}'\boldsymbol{\theta})|\mathbf{z}]$ 进行设定。一种可能性就是假定 $\mathrm{E}[\exp(\mathbf{x}'\boldsymbol{\theta}_0)\mathbf{x}|\mathbf{z}]$ 是关于 $\mathbf{z}$ 的低阶多项式，在此情况下，将存在比参数个数更多的矩条件，故通过广义矩方法进行估计，而不是用式(6.25)直接估计，同时假定误差是同方差的。若这些额外假设是错误的，则该估计量仍是一致的，给定式(6.23)是有效的，并利用式(6.16)中方差矩阵的稳健形式来获得一致标准误差。一种更普遍的方法是，直接用 $\mathbf{z}$ 而不是用 $\mathbf{D}^*(\mathbf{z},\boldsymbol{\theta}_0)$ 作为工具。

非线性回归最优矩条件事例

在一些情况下，特别是当 $\mathbf{z}=\mathbf{x}$ 时，结果(6.24)是有益的。这里将证实，GLS 是基于 $\mathrm{E}[u|\mathbf{x}]=0$ 的最有效广义矩方法估计量。

考察非线性回归模型 $y=g(\mathbf{x},\boldsymbol{\beta})+u$。如果起点是条件矩约束 $\mathrm{E}[u|\mathbf{x}]=0$ 或者 $\mathrm{E}[y-g(\mathbf{x},\boldsymbol{\beta})|\mathbf{x}]=0$，那么式(6.23)中 $\mathbf{z}=\mathbf{x}$，并且式(6.24)产生：

$$\begin{aligned}\mathbf{D}^*(\mathbf{x},\boldsymbol{\beta})&=\mathrm{E}\left[\frac{\partial}{\partial\boldsymbol{\beta}}(y-g(\mathbf{x},\boldsymbol{\beta}_0))|\mathbf{x}\right]\{\mathrm{V}[y-g(\mathbf{x},\boldsymbol{\beta}_0)|\mathbf{x}]\}^{-1}\\&=-\frac{\partial g(\mathbf{x},\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}}\times\frac{1}{\mathrm{V}[u|\mathbf{x}]}\end{aligned}$$

这仅需要对 $\mathrm{V}[u|\mathbf{x}]$ 进行设定。由式(6.25)知，最优广义矩方法估计量可直接求解对应的样本矩条件：

$$\frac{1}{N}\sum_{i=1}^{N}-\frac{\partial g(\mathbf{x}_i,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\times\frac{(y_i-g(\mathbf{x}_i,\boldsymbol{\beta}))}{\sigma_i^2}=\mathbf{0}$$

其中，$\sigma_i^2=\mathrm{V}[u_i|\mathbf{x}_i]$ 在函数形式上与 $\boldsymbol{\beta}$ 无关。当误差是异方差时，这些是广义 NLS 的一阶条件。利用 σ_i^2 的一致估计值 $\hat{\sigma}_i^2$ 来实施计算是可行的，在此情况下，广义矩方法估计与 FGNLS 是一样的。对 σ_i^2 错误设定来说，人们能获得稳健的标准误差，有关详细内容参见 5.8 节。

对线性模型 $g(\mathbf{x},\boldsymbol{\beta})=\mathbf{x}'\boldsymbol{\beta}$ 专门研究，基于 $\mathrm{E}[u|\mathbf{x}]=0$ 的最优广义矩方法估计量就是 GLS；进一步对同方差误差情况专门研究，则基于 $\mathrm{E}[u|\mathbf{x}]=0$ 的最优广义矩方法估计量正是 OLS。如同在 6.3.6 节已看到的，如果可使用额外条件的矩条件，那么进行更有效的估计是可能的。

6.3.8 对过度识别约束的检验

利用沃尔德检验(参见 5.5 节)或者由 7.5 节给出的其他方法，对 $\boldsymbol{\theta}$ 进行假设检验。

此外，存在能用于过度识别模型的十分一般的模型设定检验，使用矩条件个数(r 个)比参数个数(q 个)多的过度识别。这种检验是 $N^{-1}\sum_i\hat{\mathbf{h}}_i$ 接近于 $\mathbf{0}$ 的封闭性检验，其中，$\hat{\mathbf{h}}_i=\mathbf{h}(\mathbf{w}_i,\hat{\boldsymbol{\theta}})$。这是对 H_0：$\mathrm{E}[\mathbf{h}(\mathbf{w},\boldsymbol{\theta}_0)]=\mathbf{0}$ 进行的明显检验，即初始总体矩条件。对于恰好识别的模型来说，估计利用了 $N^{-1}\sum_i\hat{\mathbf{h}}_i=\mathbf{0}$，从而这种检验是不可行的。然而，对于过度识别模型来说，一阶条件(6.8)使 $q\times r$ 阶矩阵乘以 $N^{-1}\sum_i\hat{\mathbf{h}}_i$ 的秩为 0，其中，$q<r$，所以 $\sum_i\hat{\mathbf{h}}_i\neq\mathbf{0}$。

在特殊情况下，$\boldsymbol{\theta}$ 可以通过式(6.18)定义的 $\hat{\boldsymbol{\theta}}_{\mathrm{OGMM}}$ 得以估计，汉森(Hansen，1982)证明，在 H_0：$\mathrm{E}[\mathbf{h}(\mathbf{w},\boldsymbol{\theta}_0)]=\mathbf{0}$ 下，过度识别约束检验统计量[**overidentifying restrictions (OIR) test statistic**]：

$$\mathrm{OIR}=\left(N^{-1}\sum\nolimits_i\hat{\mathbf{h}}_i\right)'\hat{\mathbf{S}}^{-1}\left(N^{-1}\sum\nolimits_i\hat{\mathbf{h}}_i\right) \tag{6.26}$$

渐近服从 $\chi^2(r-q)$ 分布。注意到，OIR 等于 GMM 目标函数(6.18)在 $\hat{\boldsymbol{\theta}}_{\mathrm{OGMM}}$ 处的计算值。如果 OIR 很大，就要拒绝总体矩条件，而广义矩方法估计量关于 $\boldsymbol{\theta}$ 是不一致的。

式(6.26)中给出的特殊二次形式 $N^{-1}\sum_i\hat{\mathbf{h}}_i$ 在 H_0 下服从 $\chi^2(r-q)$ 分布的先验信息是不明显的。正式推导将在下一节给出，而在线性工具变量估计的情况下，其直观解释将在 8.4.4 节给出。

一种经典应用就是消费的生命周期模型(参见 6.2.7 节)，在此情况下，正交条件是欧拉条件。很大的卡方检验统计量常常表明，对生命周期假设的拒绝。然而，相反，它更应该勉强地解释为对效用函数特殊设定的拒绝，以及研究中所使用的随机假设集合。

6.3.9 广义矩方法估计量推导

通过引入更简洁记号，可简化代数运算。广义矩方法估计量极小化下式：

$$Q_N(\boldsymbol{\theta})=\mathbf{g}_N(\boldsymbol{\theta})'\mathbf{W}_N\mathbf{g}_N(\boldsymbol{\theta}) \tag{6.27}$$

其中，$\mathbf{g}_N(\boldsymbol{\theta})=N^{-1}\sum_i\mathbf{h}_i(\boldsymbol{\theta})$。那么，广义矩方法一阶条件(6.8)变成：

$$\mathbf{G}_N(\hat{\boldsymbol{\theta}})'\mathbf{W}_N\mathbf{g}_N(\hat{\boldsymbol{\theta}})=\mathbf{0} \tag{6.28}$$

其中，$\mathbf{G}_N(\boldsymbol{\theta})=\partial\mathbf{g}_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'=N^{-1}\sum_i\partial\mathbf{h}_i(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'$。

就一致性而言，考察 $\partial Q_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\boldsymbol{\theta}_0}$ 的概率极限等于 0 的非正式条件。由式

(6.28)知，如同 $\mathbf{G}_N(\boldsymbol{\theta}_0)$ 与 $\mathbf{W}_N$ 具有有限概率极限的情况一样，由命题 6.1 的假设(iii)、(iv)和(v)可知，$\text{plim}\ \mathbf{g}_N(\boldsymbol{\theta}_0)=\mathbf{0}$。更直观地讲，如果应用大数定律，同时 $\mathrm{E}[\mathbf{h}_i(\boldsymbol{\theta}_0)]=\mathbf{0}$，那么 $\mathbf{g}_N(\boldsymbol{\theta}_0)=N^{-1}\sum_i \mathbf{h}_i(\boldsymbol{\theta}_0)$ 具有概率极限 0，这是在式(6.5)开始时所做出的假设。

由重要假设(ii)以及另外的假设(iii)与(iv)知，参数 $\boldsymbol{\theta}_0$ 是可识别的，这几个假设把 $\mathbf{G}_N(\boldsymbol{\theta}_0)$ 及 $\mathbf{W}_N$ 的概率极限限制成为满秩矩阵。$\mathbf{G}_0=\text{plim}\ \mathbf{G}_N(\boldsymbol{\theta}_0)$ 是满秩矩阵的假设，称为识别(**identification**)的秩条件(**rank condition**)。识别的一个比较弱的必要条件是阶条件(**order condition**)，即 $r\geqslant q$。

对渐近正态性来说，与基于目标函数 $Q_N(\boldsymbol{\beta})=N^{-1}\sum_i q(\mathbf{w}_i,\boldsymbol{\theta})$ 的估计量相比，它需要更一般的理论，这里，$N^{-1}\sum_i q(\mathbf{w}_i,\boldsymbol{\theta})$ 刚好涉及其和式。我们通过用 $\sqrt{N}$ 乘以式(6.28)来重新标度，所以有：

$$\mathbf{G}_N(\hat{\boldsymbol{\theta}})'\mathbf{W}_N\sqrt{N}\mathbf{g}_N(\hat{\boldsymbol{\theta}})=\mathbf{0} \tag{6.29}$$

一般性定理 5.3 的方法是对式(6.28)左边在 $\boldsymbol{\theta}_0$ 处附近取泰勒级数展开式。由于 $\hat{\boldsymbol{\theta}}$ 既出现在第一项也出现在第三项中，所以这个式子很复杂，同时要求 $\mathbf{G}_N(\boldsymbol{\theta})$ 的一阶导数存在，从而要求 $\mathbf{g}_N(\boldsymbol{\theta})$ 的二阶导数存在。由于 $\mathbf{G}_N(\hat{\boldsymbol{\theta}})$ 与 $\mathbf{W}_N$ 具有有限概率，所以更直接地，仅仅对 $\sqrt{N}\mathbf{g}_N(\hat{\boldsymbol{\theta}})$ 取准确的泰勒级数表达式就足够了。这会产生类似于第 5 章曾经讨论的 m 估计表达式，满足：

$$\sqrt{N}\mathbf{g}_N(\hat{\boldsymbol{\theta}})=\sqrt{N}\mathbf{g}_N(\boldsymbol{\theta}_0)+\mathbf{G}_N(\boldsymbol{\theta}^+)\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) \tag{6.30}$$

回顾 $\mathbf{G}_N(\boldsymbol{\theta})=\partial\mathbf{g}_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'$，其中，$\boldsymbol{\theta}^+$ 表示位于 $\boldsymbol{\theta}_0$ 与 $\hat{\boldsymbol{\theta}}$ 之间的点。把式(6.30)代入式(6.29)，得到：

$$\mathbf{G}_N(\hat{\boldsymbol{\theta}})'\mathbf{W}_N[\sqrt{N}\mathbf{g}_N(\boldsymbol{\theta}_0)+\mathbf{G}_N(\boldsymbol{\theta}^+)\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)]=\mathbf{0}$$

求解 $\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$，得出：

$$\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)=-[\mathbf{G}_N(\hat{\boldsymbol{\theta}})'\mathbf{W}_N\mathbf{G}_N(\boldsymbol{\theta}^+)]^{-1}\mathbf{G}_N(\hat{\boldsymbol{\theta}})'\mathbf{W}_N\sqrt{N}\mathbf{g}_N(\boldsymbol{\theta}_0) \tag{6.31}$$

求解广义矩方法估计量的极限分布时，等式(6.31)是一个重要结果。给定一致性，即在 $\hat{\boldsymbol{\theta}}\xrightarrow{p}\boldsymbol{\theta}_0$ 条件下，可得到前五项中每一项的概率极限，在此情况下，$\boldsymbol{\theta}^+\xrightarrow{p}\boldsymbol{\theta}_0$。由假设(v)可知，式(6.31)右边最后一项具有极限正态分布。因而：

$$\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\xrightarrow{d}-(\mathbf{G}_0'\mathbf{W}_0\mathbf{G}_0)^{-1}\mathbf{G}_0'\mathbf{W}_0\times\mathcal{N}[0,\mathbf{S}_0]$$

其中，$\mathbf{G}_0$、$\mathbf{W}_0$ 以及 $\mathbf{S}_0$ 已由命题 6.1 定义。若利用极限正态乘积规则(定理 A.17)，则可得出式(6.11)。

这种推导是把广义矩方法一阶条件处理成为 r 个样本矩 $\mathbf{g}_N(\hat{\boldsymbol{\theta}})$ 的 q 个线性组合，这是因为，$\mathbf{G}_N(\hat{\boldsymbol{\theta}})'\mathbf{W}_N$ 是 $q\times r$ 阶矩阵。由于 $\mathbf{G}_N(\hat{\boldsymbol{\theta}})'\mathbf{W}_N$ 是满秩方阵，所以广义矩方法估计量是当 $q=r$ 时的特殊情况，因而 $\mathbf{G}_N(\hat{\boldsymbol{\theta}})'\mathbf{W}_N\mathbf{g}_N(\hat{\boldsymbol{\theta}})=\mathbf{0}$ 蕴含 $\mathbf{g}_N(\hat{\boldsymbol{\theta}})=\mathbf{0}$。

为了推导出式(6.26)的 OIR 检验统计量分布，以在 $\boldsymbol{\theta}_0$ 附近 $\sqrt{N}\mathbf{g}_N(\hat{\boldsymbol{\theta}})$ 的一阶泰勒级数展开开始，得出：

$$\begin{aligned}\sqrt{N}\mathbf{g}_N(\hat{\boldsymbol{\theta}}_{\mathrm{OGMM}}) &= \sqrt{N}\mathbf{g}_N(\boldsymbol{\theta}_0)+\mathbf{G}_N(\boldsymbol{\theta}^+)\sqrt{N}(\hat{\boldsymbol{\theta}}_{\mathrm{OGMM}}-\boldsymbol{\theta}_0)\\ &= \sqrt{N}\mathbf{g}_N(\boldsymbol{\theta}_0)-\mathbf{G}_0(\mathbf{G}_0'\mathbf{S}_0^{-1}\mathbf{G}_0)^{-1}\mathbf{G}_0'\mathbf{S}_0^{-1}\sqrt{N}\mathbf{g}_N(\boldsymbol{\theta}_0)+o_p(1)\\ &= [\mathbf{I}-\mathbf{M}_0\mathbf{S}_0^{-1}]\sqrt{N}\mathbf{g}_N(\boldsymbol{\theta}_0)+o_p(1)\end{aligned}$$

其中,第二个等式运用了含有对于 $\mathbf{S}_0^{-1}$ 来说一致的 $\mathbf{W}_N$ 的式(6.31),$\mathbf{M}_0=\mathbf{G}_0(\mathbf{G}_0'\mathbf{S}_0^{-1}\mathbf{G}_0)^{-1}\mathbf{G}_0'$,而 $o_p(1)$已由定义 A. 22 给出。由此可得:

$$\begin{aligned}\mathbf{S}_0^{-1/2}\sqrt{N}\mathbf{g}_N(\hat{\boldsymbol{\theta}}_{\mathrm{OGMM}}) &= \mathbf{S}_0^{-1/2}[\mathbf{I}-\mathbf{M}_0\mathbf{S}_0^{-1}]\sqrt{N}\mathbf{g}_N(\boldsymbol{\theta}_0)+o_p(1)\\ &= [\mathbf{I}-\mathbf{S}_0^{1/2}-\mathbf{M}_0\mathbf{S}_0^{-1/2}]\mathbf{S}_0^{-1/2}\sqrt{N}\mathbf{g}_N(\boldsymbol{\theta}_0)+o_p(1)\end{aligned} \tag{6.32}$$

现在,$[\mathbf{I}-\mathbf{S}_0^{-1/2}\mathbf{M}_0\mathbf{S}_0^{-1/2}]=[\mathbf{I}-\mathbf{S}_0^{-1/2}\mathbf{G}_0(\mathbf{G}_0'S_0^{-1}\mathbf{G}_0)^{-1}\mathbf{G}_0'\mathbf{S}_0^{-1/2}]$是秩为$(r-q)$的幂等矩阵,并且 $\mathbf{S}_0^{-1/2}\sqrt{N}\mathbf{g}_N(\boldsymbol{\theta}_0)\xrightarrow{d}\mathcal{N}[0,\mathbf{I}]$给出$\sqrt{N}\mathbf{g}_N(\boldsymbol{\theta}_0)\xrightarrow{d}\mathcal{N}[0,\mathbf{S}_0]$。由正态变量的二次形式标准结果可知,内积

$$\tau_N=(\mathbf{S}_0^{-1/2}\sqrt{N}\mathbf{g}_N(\hat{\boldsymbol{\theta}}_{\mathrm{OGMM}}))'(\mathbf{S}_0^{-1/2}\sqrt{N}\mathbf{g}_N(\hat{\boldsymbol{\theta}}_{\mathrm{OGMM}}))$$

收敛到 $\chi^2(r-q)$分布。

6.4 线性工具变量

回归元与误差项相关,会导致最小二乘法的非一致性。这类失效的例子包括省略变量、联立性、回归元的测量误差以及样本选择偏倚。倘若存在合适的工具,工具变量方法就能提供解决这些问题的一般方法。

当然,工具变量方法属于广义矩方法框架,因为过剩的工具导致能用于估计的矩条件过多。利用广义矩方法框架,很容易获得许多 IV 结果。

线性工具变量是非常重要的,本书许多地方都曾出现。对它的介绍已由 4.8 节与 4.9 节给出。本节阐述作为广义矩方法特殊应用的单方程线性工具变量。为了完整起见,本节还阐述一种特殊情况的较早文献,即两阶段最小二乘法估计量。系统线性工具变量估计将在 6.9.5 节加以概述。而 8.4 节则详述对线性模型的内生性检验以及对过度识别约束的检验。第 22 章将阐述具有面板数据的线性工具变量估计。

6.4.1 *带有工具的线性广义矩方法*

考察线性回归模型:

$$y_i=\mathbf{x}_i'\boldsymbol{\beta}+u_i \tag{6.33}$$

其中,若 $\mathbf{x}$ 的每个元素与模型(6.33)中的误差项相关,$\mathbf{x}$ 的每一个元素都可被看成外生回归元(**exogenous regressors**)。若所有回归元都是外生的,则使用 LS 估计量;若 $\mathbf{x}$ 的任一个元素都是内生的,则 LS 估计量关于 $\boldsymbol{\beta}$是非一致的。

由 4.8 节可知,通过工具变量估计,可获得一致估计。其关键假设是存在 $r\times1$ 维工具向量 $\mathbf{z}$(**instrument z**),它满足:

$$\mathrm{E}[u_i|\mathbf{z}_i]=\mathbf{0} \tag{6.34}$$

外生回归元能够由其自身作为工具。由于必须至少存在与回归元个数一样多的工具，所以找出至少等于模型中内生变量个数的额外工具是一种挑战。4.8.2 节已经给出这种工具的一些例子。

线性广义矩方法估计量

由 6.2.5 节可知，条件矩约束(6.34)与模型(6.33)蕴含着无条件矩约束：

$$\mathrm{E}[\mathbf{z}_i(y_i-\mathbf{x}_i'\boldsymbol{\beta})]=\mathbf{0} \tag{6.35}$$

为了令记号简洁，下述分析使用 $\boldsymbol{\beta}$ 而不是更正式的 $\boldsymbol{\beta}_0$ 来表示真实参数值。相对应的样本矩中二次形式会导致式(6.4)给出的广义矩方法目标函数 $Q_N(\boldsymbol{\beta})$。

与以往一样，使用矩阵记号定义 $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{u}$，并设 $\mathbf{Z}$ 表示 $N\times r$ 阶工具矩阵，其第 i 行为 $\mathbf{z}_i'$。那么，$\sum_i\mathbf{z}_i(y_i-\mathbf{x}_i'\boldsymbol{\beta})=\mathbf{Z}'\mathbf{u}$，式(6.4)变为：

$$Q_N(\boldsymbol{\beta})=\left[\frac{1}{N}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{Z}\right]\mathbf{W}_N\left[\frac{1}{N}\mathbf{Z}'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right] \tag{6.36}$$

其中，$\mathbf{W}_N$ 表示 $r\times r$ 阶满秩的对称加权矩阵，一个重要例子将在本节末尾给出。在这种广义矩方法特殊情况下，一阶条件：

$$\frac{\partial Q_N(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}=-2\left[\frac{1}{N}\mathbf{X}'\mathbf{Z}\right]\mathbf{W}_N\left[\frac{1}{N}\mathbf{Z}'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right]=\mathbf{0}$$

实际上能求解出 $\boldsymbol{\beta}$，导致线性工具变量模型的广义矩方法估计量(**GMM estimator in the linear IV model**)：

$$\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}=[\mathbf{X}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\mathbf{y} \tag{6.37}$$

这里，消去了被 N 除的项。

线性广义矩方法估计量分布

6.3 节的一般性结果能用于推导渐近分布。否则，由于存在 $\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}$ 的显性解，运用 4.4 节给出的 OLS 分析来适应这个要求。把 $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{u}$ 代入式(6.37)，得到：

$$\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}=\boldsymbol{\beta}+[(N^{-1}\mathbf{X}'\mathbf{Z})\mathbf{W}_N(N^{-1}\mathbf{Z}'\mathbf{X})]^{-1}(N^{-1}\mathbf{X}'\mathbf{Z})\mathbf{W}_N(N^{-1}\mathbf{Z}'\mathbf{u}) \tag{6.38}$$

由最后一项知道，广义矩方法估计量的一致性本质上要求 $\operatorname{plim} N^{-1}\mathbf{Z}'\mathbf{u}=\mathbf{0}$。在纯随机抽样条件下，需要式(6.35)成立，而在其他普通抽样方案下(参见 24.3 节)，则需要比较强的假设(6.34)。

此外，倘若 $\mathbf{W}_N$ 是满秩的，$\boldsymbol{\beta}$ 的识别秩条件(**rank condition**)即 $\operatorname{plim} N^{-1}\mathbf{Z}'\mathbf{X}$ 是秩为 K 的，这就确保了右边逆存在。一个较弱条件是 $r\geqslant K$。

极限分布建立在 $\sqrt{N}(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}-\boldsymbol{\beta})$ 表达式的基础上，该式可以通过直接对式(6.38)加以处理而得到。这就得出 $\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}$ 的渐近正态分布，其均值为 $\boldsymbol{\beta}$，而估计渐近方差为：

$$\hat{\mathrm{V}}[\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}]=N[\mathbf{X}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\mathbf{X}]^{-1}[\mathbf{X}'\mathbf{Z}\mathbf{W}_N\hat{\mathbf{S}}\mathbf{W}_N\mathbf{Z}'\mathbf{X}][\mathbf{X}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\mathbf{X}]^{-1} \tag{6.39}$$

其中，$\hat{\mathbf{S}}$ 表示

$$\mathbf{S}=\lim\frac{1}{N}\sum_{i=1}^{N}\mathrm{E}[u_i^2\mathbf{z}_i\mathbf{z}_i']$$

的一致估计值，给定对于不同 i 的具有通常横截面独立性的假设。必不可少的额外假设需要式(6.39)是 $N^{-1/2}\mathbf{Z}'\mathbf{u}\xrightarrow{d}\mathcal{N}[\mathbf{0},\mathbf{S}]$。结果(6.39)还可以由满足 $\mathbf{h}(\cdot)=\mathbf{z}(\mathbf{y}-\mathbf{x}'\boldsymbol{\beta})$ 的命题 6.1 得出，从而 $\partial\mathbf{h}/\partial\boldsymbol{\beta}'=-\mathbf{z}\mathbf{x}'$。

对带有异方差误差的横截面来说，$\mathbf{S}$ 可通过

$$\hat{\mathbf{S}}=\frac{1}{N}\sum_{i=1}^{N}\hat{u}_i^2\mathbf{z}_i\mathbf{z}_i'=\mathbf{Z}'\mathbf{D}\mathbf{Z}/N \tag{6.40}$$

一致估计出来，其中，$\hat{u}_i=y_i-\mathbf{x}_i'\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}$ 表示广义矩方法残差，而 $\mathbf{D}$ 表示 $N\times N$ 阶对角矩阵，其各个元素为 $\hat{u}_i^2$。一种广泛运用的小样本调整方法是用 $N-K$ 而不是 N 去除公式 $\hat{\mathbf{S}}$。在更有约束性的同方差误差下，$\mathrm{E}[u_i^2\mid\mathbf{z}_i]=\sigma^2$，因而 $\mathbf{S}=\lim N^{-1}\sum_i\sigma^2\mathrm{E}[\mathbf{z}_i\mathbf{z}_i']$，所以得出估计值：

$$\hat{\mathbf{S}}=s^2\mathbf{Z}'\mathbf{Z}/N \tag{6.41}$$

其中，$s^2=(N-K)^{-1}\sum_{i=1}^{N}\hat{u}_i^2$ 表示 σ^2 的一致估计值。这些结果非常类似于 4.4.5 节所述的普通最小二乘法的结果。

6.4.2 *各种不同线性广义矩方法估计量*

应用 6.4.1 节的结果，需要对加权矩阵 $\mathbf{W}_N$ 进行假定。对恰好识别模型来说，对 $\mathbf{W}_N$ 的所有选取会产生相同的估计量。对过度识别模型来说，存在对 $\mathbf{W}_N$ 的两种普遍选取方法。

当假定独立异方差误差时，表 6.2 概括了这些估计量，并给出了由式(6.39)给定的估计方差矩阵的适当设定。

表 6.2　线性工具变量模型的广义矩方法估计量及其渐近方差[a]

估计量	定义与渐近方差
GMM	$\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}=[\mathbf{X}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\mathbf{y}$
(一般 $\mathbf{W}_N$)	$\hat{\mathrm{V}}[\hat{\boldsymbol{\beta}}]=N[\mathbf{X}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\mathbf{X}]^{-1}[\mathbf{X}'\mathbf{Z}\mathbf{W}_N\hat{\mathbf{S}}\mathbf{W}_N\mathbf{Z}'\mathbf{X}][\mathbf{X}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\mathbf{X}]^{-1}$
最优 GMM	$\hat{\boldsymbol{\beta}}_{\mathrm{OGMM}}=[\mathbf{X}'\mathbf{Z}\hat{\mathbf{S}}^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}\hat{\mathbf{S}}^{-1}\mathbf{Z}'\mathbf{y}$
($\mathbf{W}_N=\hat{\mathbf{S}}^{-1}$)	$\hat{\mathrm{V}}[\hat{\boldsymbol{\beta}}]=N[\mathbf{X}'\mathbf{Z}\hat{\mathbf{S}}^{-1}\mathbf{Z}'\mathbf{X}]^{-1}$
2SLS	$\hat{\boldsymbol{\beta}}_{\mathrm{2SLS}}=[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}\mathbf{Z}'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$
($\mathbf{W}_N=[N^{-1}\mathbf{Z}'\mathbf{Z}]^{-1}$)	$\hat{\mathrm{V}}[\hat{\boldsymbol{\beta}}]=N[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\hat{\mathbf{S}}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]$ $\times[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}$
	若同方差误差，则 $\hat{\mathrm{V}}[\hat{\boldsymbol{\beta}}]=s^2[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}$
IV	$\hat{\boldsymbol{\beta}}_{\mathrm{IV}}=[\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{Z}'\mathbf{y}$
(恰好识别)	$\hat{\mathrm{V}}[\hat{\boldsymbol{\beta}}]=N(\mathbf{Z}'\mathbf{X})^{-1}\hat{\mathbf{S}}(\mathbf{X}'\mathbf{Z})^{-1}$

[a] 方程建立在线性回归模型的基础上，线性回归模型的因变量为 $\mathbf{y}$，回归元为 $\mathbf{X}$，而工具为 $\mathbf{Z}$。$\hat{\mathbf{S}}$ 已由式(6.40)定义，而 s^2 已在式(6.41)后面定义。除了对于 2SLS 估计量给定的同方差误差进行简化之外，所有的方差矩阵估计值都假定误差对于不同的观测值来说是独立且异方差的。最优广义矩方法使用最优加权矩阵。

工具变量估计量

在恰好识别 $r=K$ 的情况下，$\mathbf{X}'\mathbf{Z}$ 是一个可逆方阵。于是，$[\mathbf{X}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\mathbf{X}]^{-1}=(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{W}_N^{-1}(\mathbf{X}'\mathbf{Z})^{-1}$，从而式(6.37)简化成工具变量(**instrumental variables**)估计量：

$$\hat{\beta}_{\mathrm{IV}}=(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y} \tag{6.42}$$

这已在4.8.6节引入。对恰好识别模型来说，就 $\mathbf{W}_N$ 的任何选取而言，广义矩方法估计量都等于工具变量估计量。

简单工具变量估计量还能用于过度识别模型中，通过去掉一些工具以使该模型是恰好识别的，与利用所有工具的情况相比，这会使有效性降低。

最优加权 GMM

由6.3.5节知道，对过度识别模型来说，最有效的广义矩方法估计量，即带有最优选取加权矩阵的广义矩方法，就是把式(6.37)中的 $\mathbf{W}_N$ 设为 $\mathbf{W}_N=\hat{\mathbf{S}}^{-1}$。

线性工具变量模型的最优广义矩方法估计量或两步广义矩方法估计量(**optimal GMM estimator or two-step GMM estimator**)是：

$$\hat{\beta}_{\mathrm{OGMM}}=[(\mathbf{X}'\mathbf{Z})\hat{\mathbf{S}}^{-1}(\mathbf{Z}'\mathbf{X})]^{-1}(\mathbf{X}'\mathbf{Z})\hat{\mathbf{S}}^{-1}(\mathbf{Z}'\mathbf{y}) \tag{6.43}$$

对异方差误差来说，利用建立在第一步一致估计值 $\hat{\beta}$ 基础上的式(6.40)，比如由式(6.44)定义的2SLS估计量，就可计算 $\hat{\mathbf{S}}$。怀特(White，1982)把这种估计量称为两阶段工具变量估计量(**two-stage IV estimator**)，这是因为两步都需要工具变量估计。

表6.2给出的最优广义矩方法的估计渐近方差矩阵具有相对简单的形式，因为当 $\mathbf{W}_N=\hat{\mathbf{S}}^{-1}$ 时，式(6.39)得以简化。在计算估计方差时，人们可使用如表6.2所述的 $\hat{\mathbf{S}}$，然而，一种更普遍的方法是使用估计 $\tilde{\mathbf{S}}$，例如，也可利用式(6.40)进行计算，但需要在最优广义矩方法估计量处计算残差，而不是建立式(6.43)中 $\hat{\mathbf{S}}$ 的第一步估计值。

两阶段最小二乘法

如果误差是同方差的而不是异方差的，那么由式(6.41)知，$\hat{\mathbf{S}}^{-1}=[s^2N^{-1}\mathbf{Z}'\mathbf{Z}]^{-1}$。于是，式(6.37)可引出两阶段最小二乘法估计量(**two-stage least-squares estimator**)的 $\mathbf{W}_N=(N^{-1}\mathbf{Z}'\mathbf{Z})^{-1}$，它能以简洁形式表述成：

$$\hat{\beta}_{\mathrm{2SLS}}=[\mathbf{X}'\mathbf{P}_Z\mathbf{X}]^{-1}[\mathbf{X}'\mathbf{P}_Z\mathbf{y}] \tag{6.44}$$

其中，$\mathbf{P}_Z=\mathbf{Z}(\mathbf{Z}\mathbf{Z}')^{-1}\mathbf{Z}'$。下一节将阐述两阶段最小二乘法内容的基础。2SLS估计量还称为广义工具变量估计量[**generalized instrumental variables (GIV) estimator**]，因为它把工具变量估计量推广到工具个数比回归元个数还多的过度识别上。由于式(6.44)能以一步方式计算出来，所以它也称为一步广义矩方法(**one-step GMM**)，而最优广义矩方法则需要两步。

2SLS估计量服从渐近正态分布，其估计渐近方差已由表6.2给出。如果人们想要提防异方差误差，就应该使用一般形式；而许多引论性教科书所阐述的较简单形式，只有在误差本质上是同方差的时候才是一致的。

最优广义矩方法与 2SLS

在过度识别模型中,不但最优广义矩方法会使有效性提高,而且 2SLS 估计量也会使有效性提高。若误差是异方差,最优广义矩方法在有效性上胜过 2SLS,尽管其有效性提高并不大。7.5 节给出一些广义矩方法检验程序,而第 8 章将假定利用最优加权矩阵进行估计。与 2SLS 相比,最优广义矩方法具有需要额外计算的缺点。此外,如同 6.3.5 节讨论的,渐近理论为最优广义矩方法估计量分布提供了不好的小样本近似。

在横截面应用中,尽管其推断建立在异方差稳健标准误差的基础上,但普遍使用的却是稍欠有效的 2SLS。

更有效的广义矩方法估计

估计量 $\hat{\boldsymbol{\beta}}_{\text{XGMM}}$ 是建立在无条件矩条件 $\mathrm{E}[\mathbf{z}_i u_i]=\mathbf{0}$ 基础之上的最有效估计量,其中,$u_i=y_i-\mathbf{x}_i'\boldsymbol{\beta}$。不过,若起始点是条件矩条件 $\mathrm{E}[u_i|\mathbf{z}_i]=\mathbf{0}$,且误差是异方差的,意味着 $\mathrm{V}[u_i|\mathbf{z}_i]$ 随 $\mathbf{z}_i$ 而变化,则这就不是最佳的矩条件。

应用 6.3.7 节的一般结果,我们能把建立在 $\mathrm{E}[u_i|\mathbf{z}_i]=\mathbf{0}$ 基础上的广义矩方法估计最优矩条件写成:

$$\mathrm{E}[\mathrm{E}[\mathbf{x}_i|\mathbf{z}_i]u_i/\mathrm{V}[u_i|\mathbf{z}_i]]=\mathbf{0} \tag{6.45}$$

正如 6.3.7 节的 LS 回归例子,人们应该用误差方差 $\mathrm{V}[u|\mathbf{z}]$ 去除。不过,实施起来要比 LS 情况更困难一些,因为除了对 $\mathrm{V}[u|\mathbf{z}]$ 设定之外,还需要对 $\mathrm{E}[\mathbf{x}|\mathbf{z}]$ 模型进行设定。这可能带有额外结构。特别地,对线性联立方程组来说,$\mathrm{E}[\mathbf{x}_i|\mathbf{z}_i]$ 关于 $\mathbf{z}$ 是线性的,因而估计建立在 $\mathrm{E}[\mathbf{x}_i u_i/\mathrm{V}[u_i|\mathbf{z}_i]]=0$ 的基础上。

就线性模型而言,通常广义矩方法估计量建立在比较简单的条件 $\mathrm{E}[\mathbf{z}_i u_i]=\mathbf{0}$ 的基础上。已知这个条件,由式(6.43)定义的最优通常广义矩方法估计量是最有效的广义矩方法估计量。

6.4.3 一种可选择的两阶段最小二乘法推导

2SLS 估计量,即过度识别模型的标准 IV 估计量,已在 6.4.2 节推导出来并作为广义矩方法估计量。

这里,我们阐述 2SLS 估计量的三种其他推导。这些推导之一归功于泰尔(Theil),它提供了在推导时间上早于广义矩方法的 2SLS 最初动机。泰尔给出的解释强调了初步性处理。不过,它并不能推广到非线性模型,而广义矩方法解释则可以。

考察线性模型:

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{u} \tag{6.46}$$

它满足 $\mathrm{E}[\mathbf{u}|\mathbf{Z}]=\mathbf{0}$,而且 $\mathrm{V}[\mathbf{u}|\mathbf{Z}]=\sigma^2\mathbf{I}$。

变换模型的 GLS

用工具 $\mathbf{Z}'$ 左乘式(6.46),得到变换模型:

$$\mathbf{Z}'\mathbf{y}=\mathbf{Z}'\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}'\mathbf{u} \tag{6.47}$$

当 $r=K$ 时，这个变换模型经常用于引出工具变量估计量的动机，这是因为 $N^{-1}\mathbf{Z}'\mathbf{u}\to\mathbf{0}$，故可忽略 $\mathbf{Z}'\mathbf{u}$，从而求解出 $\hat{\boldsymbol{\beta}}=(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$。

不过，这里考察过度识别的情况。已知式(6.46)后面的假设，以 $\mathbf{Z}$ 为条件的误差 $\mathbf{Z}'\mathbf{u}$ 具有零均值，且方差为 $\sigma^2\mathbf{Z}'\mathbf{Z}$。从而，模型(6.46)中 $\boldsymbol{\beta}$ 的有效 GLS 估计量是：

$$\hat{\boldsymbol{\beta}}=[\mathbf{X}'\mathbf{Z}(\sigma^2\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\sigma^2\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} \tag{6.48}$$

这等于式(6.44)的 2SLS 估计量，因为可消去乘数 σ^2。更一般地讲，注意到，若变换模型(6.47)可由带有加权矩阵 $\mathbf{W}_N$ 的 WLS 加以估计，则可获得更一般的估计量(6.37)。

泰尔的解释

泰尔(Theil, 1953)提出，除了用在渐近形式上与误差项不相关的预测值 $\hat{\mathbf{X}}$ 代替回归元 $\mathbf{X}$ 之外，对最初模型(6.46)通过 OLS 回归进行估计。

假如在第一阶段模型(**first-stage model**)中，回归元 $\mathbf{X}$ 是工具与某个误差的线性组合，因而有：

$$\mathbf{X}=\mathbf{Z}\boldsymbol{\Pi}+\mathbf{v} \tag{6.49}$$

其中，$\boldsymbol{\Pi}$ 表示 $K\times r$ 阶矩阵。$\mathbf{X}$ 对 $\mathbf{Z}$ 的多变量 OLS 回归产生了估计量 $\hat{\boldsymbol{\Pi}}=(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ 以及 OLS 预测值 $\hat{\mathbf{X}}=\mathbf{Z}\hat{\boldsymbol{\Pi}}$，或：

$$\hat{\mathbf{X}}=\mathbf{P}_{\mathbf{Z}}\mathbf{X}$$

其中，$\mathbf{P}_{\mathbf{Z}}=\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$。$\mathbf{y}$ 对 $\hat{\mathbf{X}}$ 而不是对 $\mathbf{X}$ 进行回归，得出估计量：

$$\hat{\boldsymbol{\beta}}_{泰尔}=(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y} \tag{6.50}$$

泰尔的解释允许通过两个 OLS 回归得以计算，其中，第一阶段 OLS 给出 $\hat{\mathbf{X}}$，第二阶段给出 $\hat{\boldsymbol{\beta}}$，从而得出两阶段最小二乘法估计量(**two-stage least-squares estimator**)。

为了建立这个估计量的一致性，把线性模型(6.46)重新写成：

$$\mathbf{y}=\hat{\mathbf{X}}\boldsymbol{\beta}+(\mathbf{X}-\hat{\mathbf{X}})\boldsymbol{\beta}+\mathbf{u}$$

若回归元 $\hat{\mathbf{X}}$ 与综合误差项 $(\mathbf{X}-\hat{\mathbf{X}})\boldsymbol{\beta}+\mathbf{u}$ 是渐近不相关的，则 $\mathbf{y}$ 对 $\hat{\mathbf{X}}$ 的第二阶段 OLS 回归会产生 $\boldsymbol{\beta}$的一致估计量。若 $\hat{\mathbf{X}}$ 是一个任意代表性变量，则不存在任何理由使该式成立；不过，当 OLS 预测值正交于 OLS 残差时，$\hat{\mathbf{X}}$ 与$(\mathbf{X}-\hat{\mathbf{X}})$是不相关的。因而，$\text{plim}\,N^{-1}\hat{\mathbf{X}}'(\mathbf{X}-\hat{\mathbf{X}})\boldsymbol{\beta}=\mathbf{0}$。而且：

$$N^{-1}\hat{\mathbf{X}}'\mathbf{u}=N^{-1}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{u}=N^{-1}\mathbf{X}'\mathbf{Z}(N^{-1}\mathbf{Z}'\mathbf{Z})^{-1}N^{-1}\mathbf{Z}'\mathbf{u}$$

于是，若 $\mathbf{Z}$ 是一个有效工具，$\hat{\mathbf{X}}$ 与 $\mathbf{u}$ 是渐近不相关的，则 $\text{plim}\,N^{-1}\mathbf{Z}'\mathbf{u}=\mathbf{0}$。$\hat{\boldsymbol{\beta}}_{泰尔}$的这个一致性结果紧密地依赖于模型的线性，并且不能被推广到非线性模型。

式(6.50)中的泰尔估计量等于前面式(6.44)所定义的 2SLS 估计量。于是，有：

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{泰尔}&=(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y}\\&=(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}'\mathbf{P}_{\mathbf{Z}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{y}\\&=(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{y}\end{aligned}$$

这是 2SLS 估计量,最后等式中运用了 $\mathbf{P}_Z'\mathbf{P}_Z=\mathbf{P}_Z$。

利用泰尔方法实施 2SLS 时,需要小心谨慎。第二阶段 OLS 将给出错误的标准误差,即使误差是同方差的,因为它利用第二阶段 OLS 回归残差($\mathbf{y}-\hat{\mathbf{X}}\hat{\beta}$)而不是实际残差($\mathbf{y}-\mathbf{X}\hat{\beta}$)来估计 σ^2。在实际应用中,人们还可对异方差误差进行调整。一种更容易的方法是将 2SLS 作为选项,直接计算式(6.44)以及由表 6.2 给出的相应方差矩阵。

正如 6.5.4 节所要阐述的,这种 2SLS 解释并不总是可延续到非线性模型上。广义矩方法解释却可被推广到非线性模型,而且正因为如此,这里就更加强调它,而不是泰尔最初的对线性 2SLS 推导。

实际上,泰尔所考察的模型是,回归元 **X** 中仅仅有一部分是内生的,而其余部分是外生的。倘若 **X** 的所有外生成分都已包含在工具 **Z** 中,则前面的分析仍是可以应用的。那么,外生回归元对工具的第一阶段 OLS 回归拟合得很好,同时外生回归元的预测值等于其实际值。因此,在实际应用中,第一阶段中仅有内生变量对工具进行回归,而第二阶段是 **y** 对外生回归元与内生变量回归元的第一阶段预测值进行回归。

巴斯曼的解释

在恰好识别情况下,巴斯曼(Basmann, 1957)提出作为工具的简单工具变量估计量的 OLS 第一阶段预测值 $\hat{\mathbf{X}}=\mathbf{P}_Z\mathbf{X}$,因为确实存在与回归元 **X** 一样多的工具 $\hat{\mathbf{X}}$。这就得出:

$$\hat{\beta}_{巴斯曼}=(\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{y} \tag{6.51}$$

因为 $\text{plim}\, N^{-1}\hat{\mathbf{X}}'\mathbf{u}=\mathbf{0}$,正如对泰尔估计量所说明的,该估计量是一致的。

实际上,估计量(6.51)等于式(6.44)定义的 2SLS 估计量,这是因为 $\hat{\mathbf{X}}'=\mathbf{X}'\mathbf{P}_Z$。

这种工具变量方法将产生正确的标准误差,并能够推广到非线性背景。

6.4.4 *可供选择的标准工具变量估计量*

当一些回归元是内生的时候,6.4.2 节曾阐述的基于工具变量的最优广义矩方法与 2SLS 估计量,都是可以利用的标准估计量。切尔诺朱科夫和汉森(Chernozhukov and Hansen, 2005)阐述了分位数回归的工具变量估计量。

这里,我们简要讨论重要的可供选择的估计量,已知 4.9 节中详述的带有弱工具的 2SLS 不好的有限样本性质,这些估计量重新引起人们关注。

有限信息极大似然法

假定同方差正态误差,通过联合单方程(6.46)的极大似然估计与式(6.46)右边的内生回归元的简化式,就可获得有限信息极大似然估计量[**limited-information maximum likelihood (LIML) estimator**]。有关详细内容,可参见格林(Greene, 2003,第 402 页)或者戴维森和麦金农(Davidson and MacKinnon, 1993, 第 644~651 页)。更一般地讲,*k* 种类型估计量[例如,参见格林(Greene, 2003,第 403 页)]包括 LIML、2SLS 以及 OLS。

归功于安德森和鲁宾(Anderson and Rubin, 1949)的有限信息极大似然估计量,先于 2SLS 估计量。与 2SLS 不同,对用于联立方程组的正规化来说,有限信息极大似然估计量是不变的。另外,已知同方差误差,有限信息极大似然与 2SLS 是渐近等价的。不过,却极少使用有限信息极大似然,因为它实施起来更困难,并且与 2SLS 相比,对其解释也更难一些。贝克(Bekker, 1994)曾经阐述有限信息极大似然的小样本结果以及有限信息极大似然的推广情况,还可参见哈恩和豪斯曼(Hahn and Hausman, 2002)。

分裂样本工具变量

我们以巴斯曼把 2SLS 作为式(6.51)中给定的工具变量估计量解释作为开始。将式(6.46)中的 $\mathbf{y}$ 代入,得到:

$$\hat{\beta}=\beta+(\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{u}$$

由假设,$\text{plim}\, N^{-1}\mathbf{Z}'\mathbf{u}=\mathbf{0}$,因而 $\text{plim}\, N^{-1}\hat{\mathbf{X}}'\mathbf{u}=\mathbf{0}$,从而 $\hat{\beta}$ 是一致的。不过,由于工具变量估计的缘故,$\mathbf{X}$ 与 $\mathbf{u}$ 之间的相关意味着 $\hat{\mathbf{X}}=\mathbf{P}_Z\mathbf{X}$ 与 $\mathbf{u}$ 相关。因此,$\mathrm{E}[\hat{\mathbf{X}}'\mathbf{u}]\neq\mathbf{0}$,这会使工具变量估计量有偏。这种偏倚产生于利用 $\hat{\mathbf{X}}=\mathbf{Z}\hat{\mathbf{\Pi}}$ 而不是 $\hat{\mathbf{X}}=\mathbf{Z}\mathbf{\Pi}$ 作为工具。

可是,一种可选择的方法是使用工具预测值 $\tilde{\mathbf{X}}$,它除了满足 $\text{plim}\, N^{-1}\hat{\mathbf{X}}'\mathbf{u}=\mathbf{0}$ 之外,还具有 $\mathrm{E}[\tilde{\mathbf{X}}'\mathbf{u}]=0$ 的性质,并且使用估计量:

$$\tilde{\beta}=(\tilde{\mathbf{X}}'\mathbf{X})^{-1}\tilde{\mathbf{X}}'\mathbf{y}$$

由于 $\mathrm{E}[\tilde{\mathbf{X}}'\mathbf{u}]=\mathbf{0}$ 并不蕴含 $\mathrm{E}[(\tilde{\mathbf{X}}'\mathbf{X})^{-1}\tilde{\mathbf{X}}'\mathbf{u}]=\mathbf{0}$,这个估计量仍将是有偏的,但此偏倚可以减小。

安格里斯特和克鲁格(Angrist and Krueger, 1995)提出,通过把样本分裂成两个子样本($\mathbf{y}_1,\mathbf{X}_1,\mathbf{Z}_1$)与($\mathbf{y}_2,\mathbf{X}_2,\mathbf{Z}_2$)来获得这类工具。第一个样本用于从 $\mathbf{X}_1$ 对 $\mathbf{Z}_1$ 的回归中获得估计值 $\hat{\mathbf{\Pi}}_1$。第二个样本用于获得工具变量估计量,其中,工具 $\tilde{\mathbf{X}}_2=\mathbf{Z}_2\hat{\mathbf{\Pi}}_1$ 使用了从单独的第一个样本中所获得的 $\hat{\mathbf{\Pi}}_1$。安格里斯特和克鲁格(Angrist and Krueger, 1995)把无偏的分裂样本工具变量估计量(**unbiased split-sample IV estimator**)定义为:

$$\tilde{\beta}_{\text{USSIV}}=(\tilde{\mathbf{X}}_2'\mathbf{X}_2)^{-1}\tilde{\mathbf{X}}_2'\mathbf{y}_2$$

建立在泰尔对 2SLS 的解释基础上,分裂样本工具变量估计量(**split-sample IV estimator**)是不变的。与 2SLS 趋于 OLS 偏倚不同,这些估计量具有趋于 0 的有限样本偏倚。不过,因为仅有一半样本用于最后阶段,故损失了相当多的有效性。

刀切法工具变量

实施这种估计量的一个更有效变形,类似于仅仅通过逐一观测值来生成工具的方法。

设下标($-i$)表示去掉第 i 个观测值的运算(leave-one-out operation)。于是,对第 i 个观测值来说,我们从 $\mathbf{X}_{(-i)}$ 对 $\mathbf{Z}_{(-i)}$ 的回归中获得估计值 $\hat{\mathbf{\Pi}}_i$,并用作工具 $\tilde{\mathbf{x}}_i'=\mathbf{z}_i'\hat{\mathbf{\Pi}}_i$。当重复 N 次时,就得出第 i 行为 $\tilde{\mathbf{x}}_i'$ 的工具向量,将其记为 $\tilde{\mathbf{X}}_{(-i)}$。这就得

出刀切法 IV 估计量(**jackknife IV estimator**):

$$\widetilde{\beta}_{\mathrm{JIV}}=(\widetilde{\mathbf{X}}'_{(-i)}\mathbf{X})^{-1}\widetilde{\mathbf{X}}'_{(-i)}\mathbf{y}_2$$

这种估计量最初是由菲利普斯和黑尔(Phillips and Hale, 1977)提出的。安格里斯特、英伯斯和克鲁格(Angrist, Imbens and Krueger, 1999)以及布洛姆奎斯特和达尔伯格(Blomquist and Dahlberg, 1999)称它为刀切法估计量,因为刀切法(参见 11.5.5 节)是关于偏倚减小的省略一个的运算方法。获得第 N 个刀切法预测值 $\widetilde{\mathbf{x}}'_i$的计算要点是,利用 11.5.5 节给出的递归公式。最近两篇论文中给出的蒙特卡罗证据表明,出现偏倚减小但其方差增大的混合情况。因此,就均方误差而言,刀切法形式并不好于常规形式。比较早的菲利普斯和黑尔(Phillips and Hale, 1977)论文阐述的分析结果是,满足 $r>2(K+1)$的适度过度识别模型刀切法工具变量(JIV)估计量的有限样本偏倚小于 2SLS 的有限样本偏倚。还可参见哈恩、豪斯曼和库斯坦纳(Hahn, Hausman and Kuersteiner, 2001)。

独立加权 2SLS

与分裂样本工具变量有关的方法,是 6.3.5 节中奥尔顿吉和西格尔(Altonji and Segal, 1996)的独立加权广义矩方法估计量。把样本分裂成 G 个组,并对线性工具变量专门研究,就会产生独立的加权工具变量估计量(**independently weighted IV estimator**):

$$\widetilde{\beta}_{\mathrm{IWIV}}=\frac{1}{G}\sum_{g=1}^{G}[\mathbf{X}'_g\mathbf{Z}_g\hat{\mathbf{S}}^{-1}_{(-g)}\mathbf{Z}'_g\mathbf{X}_g]^{-1}\mathbf{X}'_g\mathbf{Z}_g\hat{\mathbf{S}}^{-1}_{(-g)}\mathbf{Z}'_g\mathbf{y}_g$$

其中,$\hat{\mathbf{S}}_{(-g)}$为利用式(6.40)定义的 $\hat{\mathbf{S}}$ 计算出来,只是去掉来自第 g 组的观测值。在面板数据的应用中,齐利亚克(Ziliak, 1997)发现,实施独立加权工具变量估计量,比实施无偏分裂样本工具变量估计量更好一些。

6.5 非线性工具变量

非线性工具变量方法,即由雨宫(Amemiya, 1974)提出的著名非线性 2SLS,允许在 NLS 估计量为非一致的——因为回归元与误差项相关——情况下,得到非线性回归模型的一致估计值。我们将这些方法阐述为对线性模型广义矩方法的直接推广。

与线性情况不同,该估计量没有显式公式,但其渐近分布可作为 6.3 节结果的一种特殊情况而获得。本节阐述单方程结果,而系统结果将在 6.10.4 节给出。一个极其重要的结果是,线性模型泰尔 2SLS 方法自然推广到非线性模型,能产生非一致参数估计值(参见 6.5.4 节)。不过,此时应使用广义矩方法。

当因变量模型是线性模型,但起因于因变量的特定性质,内生回归元的简化式是非线性的时候,就会产生一种可供选择的非线性。例如,内生回归元可以是计数的或二值结果。在那种情况下,仍然可应用前一节的线性方法。一种方法是,忽略内生回归元的特殊性质,同时实施常规线性 2SLS 或最优广义矩方法。或者,可通过适当的非线性回归,获得内生回归元拟合值,比如,如果内生回归元是计数的,那

么对所有工具进行泊松回归，然后遵循巴斯曼方法，利用这个拟合值作为计数工具，实施常规线性工具变量。尽管这两种估计量服从不同的渐近分布，但它们都是一致的。较简单的第一种方法是一种通常方法。

6.5.1 带工具的非线性广义矩方法

考察相当一般的非线性回归模型，其中，误差项可能是可加的或非可加的(参见6.2.2节)。因而，有：

$$u_i=r(y_i,\mathbf{x}_i,\boldsymbol{\beta}) \tag{6.52}$$

其中，带有可加误差的非线性模型是一种特殊情况：

$$u_i=y_i-g(\mathbf{x}_i,\boldsymbol{\beta}) \tag{6.53}$$

其中，$g(\cdot)$是一个设定函数。若$\mathrm{E}[u_i|\mathbf{x}_i]\neq 0$，则6.2.2节给出的估计量是非一致的。

假定存在r个工具$\mathbf{z}$，其中$r\geqslant K$，满足：

$$\mathrm{E}[u_i|\mathbf{z}_i]=0 \tag{6.54}$$

这与线性情况下的条件矩条件是一样的，只是$u_i=r(y_i,\mathbf{x}_i,\boldsymbol{\beta})$，而不是$u_i=y_i-\mathbf{x}_i'\boldsymbol{\beta}$。

非线性广义矩方法估计量

由期望迭代定律，式(6.54)变为：

$$\mathrm{E}[\mathbf{z}_i u_i]=\mathbf{0} \tag{6.55}$$

广义矩方法估计量就是对相应样本矩条件的二次形式求极小值。

若用矩阵记号，设$\mathbf{u}$表示$N\times 1$维误差向量，其第i个元素u_i已由式(6.52)给出；并设$\mathbf{Z}$表示$N\times r$阶工具矩阵，其第i行为$\mathbf{z}_i'$。于是，$\sum_i \mathbf{z}_i u_i=\mathbf{Z}'\mathbf{u}$，而非线性工具变量模型的广义矩方法估计量(**GMM estimator in the nonlinear IV model**) $\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}$极小化下式：

$$Q_N(\boldsymbol{\beta})=\left(\frac{1}{N}\mathbf{u}'\mathbf{Z}\right)\mathbf{W}_N\left(\frac{1}{N}\mathbf{Z}'\mathbf{u}\right) \tag{6.56}$$

其中，$\mathbf{W}_N$表示$r\times r$阶加权矩阵。与线性广义矩方法不同，一阶条件得不到$\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}$的闭形式解。

非线性广义矩方法估计量的分布

对于式(6.54)给出的$\boldsymbol{\beta}$，广义矩方法估计量是一致的，并且其渐近正态分布具有下述估计渐近方差：

$$\hat{\mathrm{V}}[\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}]=N[\hat{\mathbf{D}}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\hat{\mathbf{D}}]^{-1}[\hat{\mathbf{D}}'\mathbf{Z}\mathbf{W}_N\hat{\mathbf{S}}\mathbf{W}_N\mathbf{Z}'\hat{\mathbf{D}}][\hat{\mathbf{D}}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\hat{\mathbf{D}}]^{-1} \tag{6.57}$$

这里利用源于6.3.3节满足$\mathbf{h}(\cdot)=\mathbf{z}u$的结果，其中，$\hat{\mathbf{S}}$由下面内容给出，而$\hat{\mathbf{D}}$表示由下式定义的误差项导数的$N\times K$阶矩阵：

$$\hat{\mathbf{D}}=\left.\frac{\partial \mathbf{u}}{\partial \boldsymbol{\beta}'}\right|_{\hat{\beta}_{\mathrm{GMM}}} \tag{6.58}$$

对非可加误差来说，$\hat{\mathbf{D}}$的第i行是$\partial r(y_i,\mathbf{x}_i,\boldsymbol{\beta})/\partial\boldsymbol{\beta}'|_{\hat{\beta}}$。对于可加误差，当忽略式

(6.57)中消掉的负号时，$\hat{\mathbf{D}}$ 的第 i 行就是$\partial g(\mathbf{x}_i,\boldsymbol{\beta})/\partial\boldsymbol{\beta}'|_{\hat{\beta}}$。

对独立异方差误差来说，有：

$$\hat{\mathbf{S}} = N^{-1}\sum_i \hat{u}_i^2\mathbf{z}_i\mathbf{z}_i' \tag{6.59}$$

这类似于线性情况，只是现在$\hat{u}_i = r(y_i,\mathbf{x},\hat{\boldsymbol{\beta}})$ 或$\hat{u}_i = y_i - g(\mathbf{x},\hat{\boldsymbol{\beta}})$。

因此，非线性模型广义矩方法估计量的渐近方差与由式(6.39)给出的线性情况是相同的，变化仅为，回归元矩阵 **X** 由导数$\partial\mathbf{u}/\partial\boldsymbol{\beta}'|_{\hat{\beta}}$所代替。这与 5.8 节推导从线性到非线性最小二乘法的变化完全一样。由类似于线性工具变量推理知，用于识别的秩条件(**rank condition**)是，$\text{plim}\ N^{-1}\mathbf{Z}'\partial\mathbf{u}/\partial\boldsymbol{\beta}'|_{\beta_0}$ 的秩为 K，同时比较弱的阶条件(**order condition**)是 $r\geqslant K$。

6.5.2 *各种不同非线性广义矩方法估计量*

不同于选择加权矩阵的关于广义矩方法估计量的两种关键的专门研究方法分别是，设 $\mathbf{W}_N=\hat{\mathbf{S}}^{-1}$ 的最优广义矩方法，以及设 $\mathbf{W}_N=(\mathbf{Z}'\mathbf{Z})^{-1}$ 的非线性两阶段最小二乘法(NL2SLS)。表 6.3 概括假定独立异方差误差时，这些估计量及其有关的方差矩阵，同时给出一般 $\mathbf{W}_N$ 的结果，以及恰好识别模型的非线性工具变量结果。

表 6.3 非线性工具变量模型的广义矩方法估计量及其渐近方差[a]

估计量	定义与渐近方差
GMM	$Q_{\text{GMM}}(\boldsymbol{\beta})=\mathbf{u}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\mathbf{u}$
(一般 $\mathbf{W}_N$)	$\hat{\text{V}}[\hat{\boldsymbol{\beta}}]=N[\hat{\mathbf{D}}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\hat{\mathbf{D}}]^{-1}[\hat{\mathbf{D}}'\mathbf{Z}\mathbf{W}_N\hat{\mathbf{S}}\mathbf{W}_N\mathbf{Z}'\hat{\mathbf{D}}][\hat{\mathbf{D}}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\hat{\mathbf{D}}]^{-1}$
最优 GMM	$Q_{\text{OGMM}}(\boldsymbol{\beta})=\mathbf{u}'\mathbf{Z}\hat{\mathbf{S}}^{-1}\mathbf{Z}'\mathbf{u}$
($\mathbf{W}_N=\hat{\mathbf{S}}^{-1}$)	$\hat{\text{V}}[\hat{\boldsymbol{\beta}}]=N[\hat{\mathbf{D}}'\mathbf{Z}\hat{\mathbf{S}}^{-1}\mathbf{Z}'\hat{\mathbf{D}}]^{-1}$
NL2SLS	$Q_{\text{NL2SLS}}(\boldsymbol{\beta})=\mathbf{u}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{u}$
($\mathbf{W}_N=[N^{-1}\mathbf{Z}'\mathbf{Z}]^{-1}$)	$\hat{\text{V}}[\hat{\boldsymbol{\beta}}]=N[\hat{\mathbf{D}}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\mathbf{D}}]^{-1}[\hat{\mathbf{D}}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\hat{\mathbf{S}}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\mathbf{D}}]\times[\hat{\mathbf{D}}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\mathbf{D}}]^{-1}$
	假定同方差误差，则 $\hat{\text{V}}[\hat{\boldsymbol{\beta}}]=s^2[\hat{\mathbf{D}}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\mathbf{D}}]^{-1}$
NLIV	$\hat{\boldsymbol{\beta}}_{\text{NLIV}}$ 求解 $\mathbf{Z}'\mathbf{u}=\mathbf{0}$
(恰好识别)	$\hat{\text{V}}[\hat{\boldsymbol{\beta}}]=N(\mathbf{Z}'\hat{\mathbf{D}})^{-1}\hat{\mathbf{S}}(\hat{\mathbf{D}}'\mathbf{Z})^{-1}$

[a] 方程是具有在式(6.53)或式(6.52)中定义的误差 **u** 与工具 **Z** 的非线性回归模型。$\hat{\mathbf{D}}$ 表示误差向量对于 $\boldsymbol{\beta}'$ 的导数在 $\hat{\boldsymbol{\beta}}$ 处的计算值，而且对于具有可加误差的模型，则简化为条件均值函数关于 $\boldsymbol{\beta}'$ 的导数在 $\hat{\boldsymbol{\beta}}$ 处的计算值。$\hat{\mathbf{S}}$ 已由式(6.59)定义。除对给定的 NL2SLS 估计量进行同方差误差简化之外，所有方差矩阵估计值都假定误差对不同观测值来说是独立且异方差的。

非线性工具变量

在恰好识别的情况下，人们可直接使用对应于式(6.55)的样本矩条件。这就得出非线性工具变量模型的矩方法(**method of moment estimator in the nonlinear IV model**) $\hat{\boldsymbol{\beta}}_{\text{NLIV}}$，它是

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i u_i = \mathbf{0} \tag{6.60}$$

的解，或等价地，也是 $\mathbf{Z}'\mathbf{u}=\mathbf{0}$ 解，其渐近方差矩阵已由表 6.3 给出。

经常运用迭代法计算非线性估计量，该迭代法可获得目标函数最优值，而不是求解非线性估计方程组。对恰好识别来说，$\hat{\beta}_{\mathrm{NLIV}}$ 可被计算为极小化式(6.56)的广义矩方法估计量，该式具有对加权矩阵的任意选择，常常是 $\mathbf{W}_N=\mathbf{I}$，从而得出相同的估计值。

最优非线性广义矩方法

对过度识别模型来说，最优广义矩方法估计量使用加权矩阵 $\mathbf{W}_N=\hat{\mathbf{S}}^{-1}$。因此，非线性工具变量模型的最优广义矩方法估计量(**optimal GMM estimator in the nonlinear IV model**) $\hat{\beta}_{\mathrm{OGMM}}$ 极小化下式：

$$Q_N(\boldsymbol{\beta})=\left(\frac{1}{N}\mathbf{u}'\mathbf{Z}\right)\hat{\mathbf{S}}^{-1}\left(\frac{1}{N}\mathbf{Z}'\mathbf{u}\right) \tag{6.61}$$

由表 6.3 给出的估计渐近方差具有相对简单的形式，这是因为当 $\mathbf{W}_N=\hat{\mathbf{S}}^{-1}$ 时，式(6.57)得以简化。

如同线性情况一样，当误差是异方差时，最优广义矩方法估计量是两步估计量。在计算估计方差时，正如表 6.3 所述，人们能使用 $\hat{\mathbf{S}}$，不过一种更普遍的方法是使用估计量 $\tilde{\mathbf{S}}$，比如说，它也可利用式(6.59)进行计算，只是在计算残差时，要在最优广义矩方法估计量处而不是式(6.61)中用于建立 $\hat{\mathbf{S}}$ 的第一步估计值处加以计算。

非线性 2SLS

具有工具广义矩方法估计量的一种特殊情况是，设式(6.56)中 $\mathbf{W}_N=(N^{-1}\mathbf{Z}'\mathbf{Z})^{-1}$。这就得出非线性两阶段最小二乘法(**nonlinear two-stage least-squares**)估计量 $\hat{\beta}_{\mathrm{NL2SLS}}$，它极小化下式：

$$Q_N(\boldsymbol{\beta})=\frac{1}{N}\mathbf{u}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{u} \tag{6.62}$$

该估计量因作为最优广义矩方法估计量而引人注目，若误差是同方差的，则 $\hat{\mathbf{S}}=s^2\mathbf{Z}'\mathbf{Z}/N$，其中，$s^2$ 表示常值 $\mathrm{V}[u|\mathbf{z}]$ 的一致估计值，从而 $\hat{\mathbf{S}}^{-1}$ 是 $(\mathbf{Z}'\mathbf{Z})^{-1}$ 的倍数。

对同方差误差来说，该估计量具有较简单的估计渐近方差，如表 6.3 所示，这是一些教科书经常给出的结果。不过，在微观经济计量学的应用中，一种普遍的做法是允许异方差误差，并使用由表 6.3 给出的更复杂的稳健估计值。

由雨宫(Amemiya，1974)提出的 NL2SLS 估计量，是广义矩方法的重要前身。该估计量提出的动机类似于 6.4.3 节给出 2SLS 的第一个动机。因而，用工具 $\mathbf{Z}'$ 左乘模型误差 $\mathbf{u}$，得到 $\mathbf{Z}'\mathbf{u}$，其中，由于 $\mathrm{E}[\mathbf{u}|\mathbf{Z}]=\mathbf{0}$，故 $\mathrm{E}[\mathbf{Z}'\mathbf{u}]=\mathbf{0}$。于是，实施非线性 GLS 回归。假定同方差误差，这将对下式求极小值：

$$Q_N(\boldsymbol{\beta})=\mathbf{u}'\mathbf{Z}[\sigma^2\mathbf{Z}'\mathbf{Z}]^{-1}\mathbf{Z}'\mathbf{u}$$

因为 $\mathrm{V}[\mathbf{u}|\mathbf{Z}]=\sigma^2\mathbf{I}$ 蕴含 $\mathrm{V}[\mathbf{Z}'\mathbf{u}|\mathbf{Z}]=\sigma^2\mathbf{Z}'\mathbf{Z}$。此目标函数刚好是式(6.62)的一个数量倍数。

泰尔对线性 2SLS 两阶段的解释，并不总是被推广到非线性模型上（参见 6.5.4 节）。另外，很明显，NL2SLS 是一个一步估计量。雨宫选用 NL2SLS 这个名字，是因为它允许利用工具变量进行一致估计，就如同线性情况一样。该名称不应按字面意义理解，表述比较清楚的术语是非线性工具变量（**nonlinear IV**）或非线性广义工具变量估计（**nonlinear generalized IV estimation**）。

非线性模型工具选择

前面的估计量都假定诸如 $\mathrm{E}[u|\mathbf{z}]=\mathbf{0}$ 的工具存在，且建立在无条件矩条件 $\mathrm{E}[\mathbf{z}u]=\mathbf{0}$ 的基础上，则估计是最佳的。

考察具有可加误差的非线性模型，从而 $u=y-g(\mathbf{x},\boldsymbol{\beta})$。为了使工具适宜，工具必须与回归元 $\mathbf{x}$ 是相关的；不过，为了使工具有效，它不能直接作为 y 的因果变量。由式(6.57)给出的方差矩阵可知，该工具确实与具有$\partial g/\partial\boldsymbol{\beta}$的 $\mathbf{z}$ 相关，而不是与起作用的 $\mathbf{x}$ 相关，以此确保 $\hat{\mathbf{D}}'\mathbf{Z}$ 充分大。像 4.9 节研究的线性情况一样，弱工具关注的正是这里所述的有关内容。

给定 $\mathrm{E}[u|\mathbf{z}]=\mathbf{0}$，已知可能的异方差性，估计建立在那种可能不使 $\mathrm{E}[\mathbf{z}u]=\mathbf{0}$ 的最优矩条件基础上。不过，由 6.3.7 节知，最优矩条件需要难以做出的其他矩假设，因此，一种标准做法是如同这里所做的，使用 $\mathrm{E}[\mathbf{z}u]=\mathbf{0}$。

一种可供选择的控制异方差性方法是，将广义矩方法估计建立在定义成接近于同方差的误差项的基础上。例如，就计数数据而言，不是使用 $u=y-\exp(\mathbf{x}'\boldsymbol{\beta})$，而是运用标准化的误差 $u^*=u/\sqrt{\exp(\mathbf{x}'\boldsymbol{\beta})}$（参见 6.2.2 节）。然而，要注意到，$\mathrm{E}[u^*|\mathbf{z}]=0$ 与 $\mathrm{E}[u|\mathbf{z}]=0$ 是不同假设。

通常，仅有 $\mathbf{x}$ 的一个元素与 u 相关。那么，如同线性情况一样，把外生元素用作它们自身的工具，并且其挑战是找出与 u 不相关的另外工具。一些非线性应用源于如同 6.2.7 节的正式经济模型，在那种情况下，可利用信息集的许多子元素作为工具。

6.5.3 *泊松工具变量例子*

具有外生回归元的泊松回归模型，设定 $\mathrm{E}[y|\mathbf{x}]=\exp(\mathbf{x}'\boldsymbol{\beta})$。可将此看作具有可加误差 $u=y-\exp(\mathbf{x}'\boldsymbol{\beta})$的模型。若回归元是内生的，则 $\mathrm{E}[u|\mathbf{x}]\neq 0$，而且泊松 MLE 将是不一致的。一致估计要假定满足 $\mathrm{E}[u|\mathbf{z}]=0$ 的工具 $\mathbf{z}$ 存在，或者等价地：

$$\mathrm{E}[y-\exp(\mathbf{x}'\boldsymbol{\beta})|\mathbf{z}]=0$$

可直接应用前面的一些结果。其目标函数是：

$$Q_N(\boldsymbol{\beta})=\left[N^{-1}\sum\nolimits_i \mathbf{z}_i u_i\right]'\mathbf{W}_N\left[N^{-1}\sum\nolimits_i \mathbf{z}_i u_i\right]$$

其中，$u_i=y_i-\exp(\mathbf{x}_i'\boldsymbol{\beta})$。于是，其一阶条件是：

$$\left[\sum\nolimits_i \exp(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i\mathbf{z}_i'\right]\mathbf{W}_N\left[\sum\nolimits_i \mathbf{z}_i(y_i-\exp(\mathbf{x}_i'\boldsymbol{\beta}))\right]=\mathbf{0}$$

渐近分布已由表 6.3 给出，满足 $\hat{\mathbf{D}}'\mathbf{Z}=\sum_i e^{\mathbf{x}_i'\hat{\boldsymbol{\beta}}}\mathbf{x}_i\mathbf{z}_i'$，因为$\partial g/\partial\boldsymbol{\beta}=\exp(\mathbf{x}'\boldsymbol{\beta})\mathbf{x}$，并且 $\hat{\mathbf{S}}$ 已由式(6.39)定义，$\hat{u}_i=y_i-\exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}})$。最优广义矩方法估计量与 NL2SLS 估计

量在加权矩阵是 $\hat{\mathbf{S}}^{-1}$ 还是 $(N^{-1}\mathbf{Z}'\mathbf{Z})^{-1}$ 的选取方面各不相同，其中，$\mathbf{Z}'\mathbf{Z}=\sum_i \mathbf{z}_i\mathbf{z}_i'$。

一种可供选择的一致估计量由巴斯曼方法得出。首先，已知 K 个预测值 $\hat{\mathbf{x}}_i=\hat{\mathbf{\Pi}}\mathbf{z}_i$，通过 OLS 估计出简化式 $\mathbf{x}_i=\mathbf{\Pi}'\mathbf{z}_i+\mathbf{v}_i$。其次，通过如同式(6.60)的非线性工具变量，使用 $\hat{\mathbf{x}}_i$ 而不是 $\mathbf{z}_i$ 进行估计。已知 $\hat{\mathbf{\Pi}}$ 的 OLS 公式，这个估计量是下式的解：

$$\left[\sum_i \mathbf{x}_i\mathbf{z}_i'\right]\left[\sum_i \mathbf{z}_i\mathbf{z}_i'\right]^{-1}\left[\sum_i \mathbf{z}_i(y_i-\exp(\mathbf{x}_i'\boldsymbol{\beta}))\right]=\mathbf{0}$$

该估计量不同于 NL2SLS，原因在于左边第一项不同。对于线性模型来说，推广泰尔方法的潜在问题将在下一节详述。

除泊松回归之外，类似问题还会出现在非线性模型譬如二值数据模型中。

6.5.4 非线性模型两阶段估计

在非线性模型中，对线性 2SLS 进行的通常解释将会失效。因此，假定 y 具有均值 $g(\mathbf{x},\boldsymbol{\beta})$，并存在回归元 $\mathbf{x}$ 的工具 $\mathbf{z}$。那么，为了获得拟合值 $\hat{\mathbf{x}}$，如同现在所要阐明的，在 y 对 $g(\hat{\mathbf{x}},\boldsymbol{\beta})$ 进行 NLS 回归之后，要实施 $\mathbf{z}$ 对工具 $\mathbf{z}$ 的 OLS 回归，这就得出 $\boldsymbol{\beta}$ 的非一致参数估计值。不过，人们需要使用前一节阐述的 NL2SLS 估计量。

考察下述简单模型，它是建立在雨宫(Amemiya，1984)所阐述模型的基础上，也就是说，尽管模型关于参数是线性的，但关于变量则是非线性的。设：

$$y=\beta x^2+u \tag{6.63}$$
$$x=\pi z+v$$

其中，零均值误差 u 与 v 是相关的。回归元 x^2 是内生的，这是因为，x 是 v 的函数，且由假设可知，u 与 v 是相关的。因此，β 的 OLS 估计量是非一致的。假如 z 是由模型中的其他随机变量独立生成的，则 z 是一个有效工具，因为显然它与 u 独立，但与 x 相关。

工具变量估计量是 $\hat{\beta}_{\mathrm{IV}}=(\sum_i z_i x_i^2)^{-1}\sum_i z_i y_i$。这可以通过运用工具 z，进行常规的 y 对 x^2 的工具变量回归来实施。正如人们所料，经过一些代数运算之后，$\hat{\beta}_{\mathrm{IV}}$ 等于式(6.60)所定义的非线性工具变量估计量。

然而，假设我们进行下述两阶段最小二乘法估计。首先，为了得到 $\hat{x}=\hat{\pi}z$，要实施 x 对 z 回归，然后实施 y 对 $\hat{x}^2$ 回归。于是，$\hat{\beta}_{\mathrm{2SLS}}=(\sum_i \hat{x}_i^2\hat{x}_i^2)^{-1}\sum_i \hat{x}_i^2 y_i$，其中，$\hat{x}_i^2$ 表示由 x 对 z 的 OLS 回归所得到的预测值 $\hat{x}_i$ 的平方。这就得出非一致估计量。对 6.4.3 节的线性情况加以改进，我们有：

$$\begin{aligned} y_i &= \beta x_i^2+u_i \\ &= \beta\hat{x}_i^2+w_i \end{aligned}$$

其中，$w_i=\beta(x_i^2-\hat{x}_i^2)+u_i$。$y_i$ 对 $\hat{x}_i^2$ 的 OLS 回归关于 β 是非一致的，因为回归元 $\hat{x}_i^2$ 与综合误差项 w_i 是渐近相关的。正式地讲，利用 $\mathrm{plim}\,\hat{\pi}=\pi$ 并进行一些代数运算，尽管 z_i 与 v_i 独立，但是，$(x_i^2-\hat{x}_i^2)=(\pi z_i+v_i)^2-(\hat{\pi}z_i)^2=\pi^2z_i^2+2\pi z_iv_i+v_i^2-\hat{\pi}^2z_i^2$ 蕴含着 $\mathrm{plim}\,N^{-1}\sum_i\hat{x}_i^2(x_i^2-\hat{x}_i^2)=\mathrm{plim}\,N^{-1}\sum_i\pi^2z_i^2v_i^2\neq 0$。因此，$\mathrm{plim}\,N^{-1}\times$

$\sum_i \hat{x}_i^2 w_i = \text{plim}\, N^{-1}\sum_i \hat{x}_i^2\beta(x_i-\hat{x}_i)^2 \neq 0$。

不过，作为一致估计的一种变形是，在第一阶段，要求 x^2 对 z 回归而不是 x 对 z 回归，并且在第二阶段，使用预测值 $\widehat{x^2} \neq (\hat{x})^2$。可以证明，这等于 $\hat{\beta}_{\text{IV}}$。这里需要 x^2 的工具成为 x^2 的拟合值，而不是 x 拟合值的平方。

这个例子可被推广到其他非线性模型，其中，非线性是仅仅关于回归元的，因而有：

$$y=\mathbf{g}(\mathbf{x})'\boldsymbol{\beta}+u$$

其中，$\mathbf{g}(\mathbf{x})$表示 $\mathbf{x}$ 的非线性函数。一个普遍的例子就是运用幂与自然对数。假定 $\mathrm{E}[u|\mathbf{z}]=0$。为了得到预测值 $\hat{\mathbf{x}}$，可通过 $\mathbf{x}$ 对 $\mathbf{z}$ 回归而获得非一致估计值，然后求 y 对 $\mathbf{g}(\hat{\mathbf{x}})$的回归。为了得到预测值 $\hat{\mathbf{g}}(\mathbf{x})$，可通过 $\mathbf{g}(\mathbf{x})$对 $\mathbf{z}$ 回归获得一致估计值，然后第二阶段要求 y 对 $\hat{\mathbf{g}}(\mathbf{x})$的回归。我们使用 $\hat{\mathbf{g}}(\mathbf{x})$而不是 $\mathbf{g}(\hat{\mathbf{x}})$作为 $\mathbf{g}(\mathbf{x})$的工具。于是，即使第二阶段回归给出无效标准误差，但是，OLS 将使用残差$\hat{u}=y-\hat{\mathbf{g}}(\mathbf{x})'\hat{\boldsymbol{\beta}}$而不是$\hat{u}=y-\mathbf{g}(\mathbf{x})'\hat{\boldsymbol{\beta}}$。一种最佳方法是，直接使用广义矩方法或 NL2SLS 命令。

更一般地讲，模型可能关于变量和参数都是非线性的。考虑具有可加误差的单指标模型，因而有：

$$y=g(\mathbf{x}'\boldsymbol{\beta})+u$$

为了获得预测值 $\hat{\mathbf{x}}$，通过 $\mathbf{x}$ 对 $\mathbf{z}$ 的 OLS，得到非一致估计值，然后要求 y 对 $g(\hat{\mathbf{x}}'\boldsymbol{\beta})$ 的 NLS 回归。这里需要使用广义矩方法，或者使用 NL2SLS。从本质上看，为了一致性，我们需要 $\hat{g}(\mathbf{x}'\boldsymbol{\beta})$，而不是 $g(\hat{\mathbf{x}}'\boldsymbol{\beta})$。

NL2SLS 例子

我们考察具有简单非线性模型的 NL2SLS 估计，该非线性由内生变量的平方作为回归元引起，如同前一节一样。

由于数据生成过程是式(6.63)，所以 $y=\beta x^2+u$ 且 $x=\pi z+v$，其中，对于所有观测值，$\beta=1$，$\pi=1$，$z=1$，同时$(u,\ v)$服从联合正态分布，其均值为 0、方差为 1，且相关系数为 0.8。抽取的样本量为 200。其结果如表 6.4 所示。

表 6.4 非线性两阶段最小二乘法例子[a]

变量	估计量		
	OLS	NL2SLS	两阶段
x^2	1.189	0.969	1.642
	(0.025)	(0.041)	(0.172)
R^2	0.88	0.84	0.80

[a] 下一节给出的数据生成过程具有等于 1 的实际参数。该样本量 $N=200$。

这里的非线性是相当弱的，原因在于，是 x 的平方而不是 x 作为回归元。所关注的内容是对 x^2 系数 β 的估计。OLS 估计量是非一致的，而 NL2SLS 却是一致的。两阶段方法，即第一阶段要求 x 对 z 的 OLS 回归，进而得出$\hat{x}$，然后求 y 对$(\hat{x})^2$ 的 OLS 回归，这就得出下面估计值，该值偏离 $\beta=1$ 的真实值超过了两个标准误差。模拟研究表明，拟合优度有些损失，且预测值具有较大的标准误差，可是 R^2 却较小，这一点类似于线性工具变量。

6.6 时序两步 m 估计

时序两步估计方法最终关注的参数估计值是建立在未知参数最初估计的基础上。当误差具有条件方差 $\exp(\mathbf{z}'\boldsymbol{\gamma})$时,一个例子就是可行 GLS。已知 $\boldsymbol{\gamma}$ 的估计值$\tilde{\boldsymbol{\gamma}}$,FGLS 估计量$\hat{\boldsymbol{\beta}}$是 $\sum_{i=1}^{N}(y_i-\mathbf{x}_i'\hat{\boldsymbol{\beta}})/\exp(\mathbf{z}_i'\tilde{\boldsymbol{\gamma}})$ 的解。第二个例子是将在 16.10.2 节给出的赫克曼两步估计量。

这些估计量深受人们喜爱,因为它们能提供一种相对简单的方法来获得一致参数估计值。不过,为了实施有效统计推断,必须对第二步估计量的渐近方差进行调整,以便考虑到第一步估计。我们阐述特殊情况的一些结果,即第一步估计量的估计方程对样本平均值设定为 0,而第二步估计量的估计方程也将样本平均值设定为 0,这正是 m 估计量、矩方法以及估计方程估计量的情况。

把参数向量 $\boldsymbol{\theta}$ 分割成$\boldsymbol{\theta}_1$ 与 $\boldsymbol{\theta}_2$ 两部分,而最终的关注内容是 $\boldsymbol{\theta}_2$。该模型可首先求解 $\sum_{i=1}^{N}\mathbf{h}_{1i}(\hat{\boldsymbol{\theta}}_1)=\mathbf{0}$ 来获得 $\hat{\boldsymbol{\theta}}_1$,然后已知 $\hat{\boldsymbol{\theta}}_1$,求解 $N^{-1}\sum_{i=1}^{N}\mathbf{h}_{1i}(\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_2)=\mathbf{0}$,从而获得 $\hat{\boldsymbol{\theta}}_2$。通常,给定估计 $\hat{\boldsymbol{\theta}}_1$ 时 $\hat{\boldsymbol{\theta}}_2$ 的分布,不同于当 $\boldsymbol{\theta}_1$已知时 $\hat{\boldsymbol{\theta}}_2$ 的分布,而且前者比后者更复杂。除在本节末尾给出的某些特殊情况以外,如果不能考虑这种复杂情况,那么统计推断就是无效的。

下述推导由纽韦(Newey, 1984)给出,而墨菲和托佩尔(Murphy and Topel, 1985)以及帕甘(Pagan,1986)也得到了相似结果。两步估计量能重新写成一步估计量,其中,($\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$)联合求解方程:

$$N^{-1}\sum_{i=1}^{N}\mathbf{h}_1(\mathbf{w}_i,\hat{\boldsymbol{\theta}}_1)=\mathbf{0} \tag{6.64}$$

$$N^{-1}\sum_{i=1}^{N}\mathbf{h}_2(\mathbf{w}_i,\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_2)=\mathbf{0}$$

若定义 $\boldsymbol{\theta}=(\boldsymbol{\theta}_1' \quad \boldsymbol{\theta}_2')'$且 $\mathbf{h}_i=(\mathbf{h}_{i1}' \quad \mathbf{h}_{i2}')'$,将该方程写成:

$$N^{-1}\sum_{i=1}^{N}\mathbf{h}(\mathbf{w}_i,\hat{\boldsymbol{\theta}})=\mathbf{0}$$

在这种背景下,假定 $\dim(\mathbf{h}_1)=\dim(\boldsymbol{\theta}_1)$且 $\dim(\mathbf{h}_2)=\dim(\boldsymbol{\theta}_2)$,则估计方程个数等于参数个数。那么,式(6.64)是估计方程估计量或者矩方法估计量。

一致性要求 $\operatorname{plim} N^{-1}\sum_i \mathbf{h}(\mathbf{w}_i,\boldsymbol{\theta}_0)=\mathbf{0}$,其中,$\boldsymbol{\theta}_0=[\boldsymbol{\theta}_{10}', \boldsymbol{\theta}_{20}']$。如果在第一步中,$\hat{\boldsymbol{\theta}}_1$ 关于 $\boldsymbol{\theta}_{10}$是一致的,并且如果已知 $\boldsymbol{\theta}_{10}$(不是由 $\hat{\boldsymbol{\theta}}_1$ 估计的)时的 $\hat{\boldsymbol{\theta}}_2$ 第二步估计可以产生 $\boldsymbol{\theta}_{20}$的一致估计值,则这个条件应该得以满足。在矩方法框架下,要求 $\mathrm{E}[\mathbf{h}_{1i}(\boldsymbol{\theta}_1)]=\mathbf{0}$ 且 $\mathrm{E}[\mathbf{h}_{2i}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)]=\mathbf{0}$。这里假定可以建立一致性。

为了得到渐近分布,我们应用一般性结果,即:

$$\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\xrightarrow{d}\mathcal{N}[\mathbf{0},\mathbf{G}_0^{-1}\mathbf{S}_0(\mathbf{G}_0^{-1})']$$

其中,$\mathbf{G}_0$ 与 $\mathbf{S}_0$ 已由命题 6.1 定义。以类似于分割 $\boldsymbol{\theta}$ 与 $\mathbf{h}_i$ 的方式,分割 $\mathbf{G}_0$ 与 $\mathbf{S}_0$。

于是，利用$\partial \mathbf{h}_{1i}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}_2=\mathbf{0}$，得出：

$$\mathbf{G}_0 = \lim \frac{1}{N}\sum_{i=1}^{N}\mathrm{E}\begin{bmatrix}\partial \mathbf{h}_{1i}/\partial \boldsymbol{\theta}_1' & \mathbf{0}\\ \partial \mathbf{h}_{2i}/\partial \boldsymbol{\theta}_1' & \partial \mathbf{h}_{2i}/\partial \boldsymbol{\theta}_2'\end{bmatrix}=\begin{bmatrix}\mathbf{G}_{11} & \mathbf{0}\\ \mathbf{G}_{21} & \mathbf{G}_{22}\end{bmatrix}$$

这是因为由式(6.64)可知，$\mathbf{h}_{1i}(\boldsymbol{\theta})$不是$\boldsymbol{\theta}_2$的函数。由于$\mathbf{G}_0$、$\mathbf{G}_{11}$和$\mathbf{G}_{22}$都是方阵，有：

$$\mathbf{G}_0^{-1}=\begin{bmatrix}\mathbf{G}_{11}^{-1} & \mathbf{0}\\ -\mathbf{G}_{22}^{-1}\mathbf{G}_{21}\mathbf{G}_{11}^{-1} & \mathbf{G}_{22}^{-1}\end{bmatrix}$$

显然，有：

$$\mathbf{S}_0 = \lim \frac{1}{N}\sum_{i=1}^{N}\mathrm{E}\begin{bmatrix}\mathbf{h}_{1i}\mathbf{h}_{1i}' & \mathbf{h}_{1i}\mathbf{h}_{2i}'\\ \mathbf{h}_{2i}\mathbf{h}_{1i}' & \mathbf{h}_{2i}\mathbf{h}_{2i}'\end{bmatrix}=\begin{bmatrix}\mathbf{S}_{11} & \mathbf{S}_{12}\\ \mathbf{S}_{21} & \mathbf{S}_{22}\end{bmatrix}$$

$\hat{\boldsymbol{\theta}}_2$ 的渐近方差是 $\hat{\boldsymbol{\theta}}$ 方差矩阵的一个(2，2)子矩阵。经过一些代表运算后，得到：

$$\mathrm{V}[\hat{\boldsymbol{\theta}}_2]=\mathbf{G}_{22}^{-1}\begin{Bmatrix}\mathbf{S}_{22}+\mathbf{G}_{21}[\mathbf{G}_{11}^{-1}\mathbf{S}_{11}\mathbf{G}_{11}^{-1}]\mathbf{G}_{21}'\\ -\mathbf{G}_{21}\mathbf{G}_{11}^{-1}\mathbf{S}_{12}-\mathbf{S}_{21}\mathbf{G}_{11}^{-1}\mathbf{G}_{21}'\end{Bmatrix}\mathbf{G}_{22}^{-1} \tag{6.65}$$

通常计算机输出会产生不正确的标准误差，同时低估了真实标准误差，这是因为假定 $\mathrm{V}[\hat{\boldsymbol{\theta}}_2]$为 $\mathbf{G}_{22}^{-1}\mathbf{S}_{22}\mathbf{G}_{22}^{-1}$，可以证明，它小于式(6.65)给出的真实方差。

在 $\mathrm{E}[\partial \mathbf{h}_{2i}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}_1]=\mathbf{0}$ 的特殊情况下，第一步估计所引起的第二步的额外变异性，是因为 $\mathbf{G}_{21}=\mathbf{0}$ 与式(6.65)的 $\mathrm{V}[\hat{\boldsymbol{\theta}}_2]$会简化成 $\mathbf{G}_{22}^{-1}\mathbf{S}_{22}\mathbf{G}_{22}^{-1}$。

$\mathbf{G}_{21}=\mathbf{0}$ 的一个著名例子是 FGLS。那么，对异方差性来说，有：

$$\mathbf{h}_{2i}(\boldsymbol{\theta})=\frac{\mathbf{x}_{2i}(y_i-\mathbf{x}_i'\boldsymbol{\theta}_2)}{\sigma(\mathbf{x}_i,\boldsymbol{\theta}_1)}$$

其中，$\mathrm{V}[y_i|\mathbf{x}_i]=\sigma^2(\mathbf{x}_i,\boldsymbol{\theta}_1)$，并且：

$$\mathrm{E}[\partial \mathbf{h}_{2i}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}_1]=\mathrm{E}\left[-\mathbf{x}_{2i}\frac{(y_i-\mathbf{x}_i'\boldsymbol{\theta}_2)}{\sigma(\mathbf{x}_i,\boldsymbol{\theta}_1)^2}\frac{\partial\sigma(\mathbf{x}_i,\boldsymbol{\theta}_1)}{\partial\boldsymbol{\theta}_1}\right]$$

上式等于 0，因为 $\mathrm{E}[y_i|\mathbf{x}_i]=\mathbf{x}_i'\boldsymbol{\theta}_2$。进一步地，对 FGLS 来说，$\hat{\boldsymbol{\theta}}_2$ 的一致性并不要求 $\hat{\boldsymbol{\theta}}_1$ 是一致的，因为 $\mathrm{E}[\mathbf{h}_{2i}(\boldsymbol{\theta})]=\mathbf{0}$ 只需要 $\mathrm{E}[y_i|\mathbf{x}_i]=\mathbf{x}_i'\boldsymbol{\theta}_2$，而这并不依赖于 $\boldsymbol{\theta}_1$。

$\mathbf{G}_{21}=\mathbf{0}$ 的第二个例子是具有分块对角矩阵的 ML 估计，因而 $\mathrm{E}[\partial^2\mathcal{L}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}_1\partial\boldsymbol{\theta}_2']=\mathbf{0}$。这是正态性条件下回归例子的情况，其中，$\boldsymbol{\theta}_1$表示方差参数，而 $\boldsymbol{\theta}_2$表示回归参数。

不过，在其他一些例子中，$\mathbf{G}_{21}\neq\mathbf{0}$，且需要使用更繁琐的表达式(6.65)。对于某些标准的两步估计量来说，譬如由 16.5.4 节给出的著名的样本选择模型赫克曼两步估计量，都是通过计算机软件包自动实施。否则，需要对 $\mathrm{V}[\hat{\boldsymbol{\theta}}_2]$进行手工计算。其许多元素源自前面的估计。特别地，$\mathbf{G}_{11}^{-1}\mathbf{S}_{11}\mathbf{G}_{11}^{-1}$ 是 $\hat{\boldsymbol{\theta}}_1$ 的稳健方差矩阵，而 $\mathbf{G}_{22}^{-1}\mathbf{S}_{22}\mathbf{G}_{22}^{-1}$ 是 $\hat{\boldsymbol{\theta}}_2$ 的稳健方差矩阵估计值，这错误地忽略了 $\hat{\boldsymbol{\theta}}_1$ 的估计误差。对于不同 i 的数据独立而言，$\mathbf{S}_0$ 子矩阵的子成分可由 $\hat{\mathbf{S}}_{jk}=N^{-1}\sum_i\hat{\mathbf{h}}_{ji}\hat{\mathbf{h}}_{ki}'$，$j,k=1,2$ 来一致估计。这就导致了对 $\hat{\mathbf{G}}_{2i}=N^{-1}\sum_i\partial\mathbf{h}_{2i}/\partial\boldsymbol{\theta}_1'|_{\hat{\theta}}$进行计算的重要挑战。

假定使用广义矩方法，则一个备受推荐的比较简单的方法是求自助标准误差（参见10.2.5节），或者直接联合估计组合模型(6.64)中的$\boldsymbol{\theta}_1$与$\boldsymbol{\theta}_2$。

这些较简单的方法还能用于时序估计量，它们是广义矩方法估计量而不是m估计量。于是，把这两种估计量结合起来，将会产生比式(6.64)更为复杂的一系列集合，从而不能再次得到式(6.65)。不过，人们仍能进行自助法或者联合估计，而不是采用时序形式。

6.7 最小距离估计

最小距离估计提供了一种估计结构参数$\boldsymbol{\theta}$的方法，这里，$\boldsymbol{\theta}$是已知$\boldsymbol{\pi}$的一致估计值$\hat{\boldsymbol{\pi}}$时，对简化式参数$\boldsymbol{\pi}$的设定函数。

一个标准参考文献是弗格森(Ferguson, 1958)。罗腾伯格(Rothenberg, 1973)将这一方法应用到线性联立方程模型上，尽管由6.9.6节给出的一种可供选择的方法是人们运用的标准方法。最小距离估计最常用于面板数据分析之中。最初，在张伯伦(Chamberlain, 1982, 1984)所做的研究工作中（参见22.2.7节），他令$\hat{\boldsymbol{\pi}}$表示来自当前时期因变量对所有时期回归元的线性回归的OLS估计值。后来的应用则针对协方差结构（参见22.5.4节），设$\hat{\boldsymbol{\pi}}$表示面板数据的估计方差与自协方差。而且，可参见间接推断方法（12.6节）。

假定q个结构参数与$r>q$个简化式参数之间的关系是$\boldsymbol{\pi}_0=\mathbf{g}(\boldsymbol{\theta}_0)$。进一步地，假定我们具有简化式参数的一致估计值$\hat{\boldsymbol{\pi}}$。一个明显的估计量是使得$\hat{\boldsymbol{\pi}}=\mathbf{g}(\hat{\boldsymbol{\theta}})$的$\hat{\boldsymbol{\theta}}$，但由于$q<r$，这是不可行的。相反，最小距离估计量[**minimum distance (MD) estimator**]$\hat{\boldsymbol{\theta}}_{\mathrm{MD}}$是对于$\boldsymbol{\theta}$，对目标函数

$$Q_N(\boldsymbol{\theta})=(\hat{\boldsymbol{\pi}}-\mathbf{g}(\boldsymbol{\theta}))'\mathbf{W}_N(\hat{\boldsymbol{\pi}}-\mathbf{g}(\boldsymbol{\theta})) \tag{6.66}$$

求极小值，其中，$\mathbf{W}_N$表示$r\times r$阶加权矩阵。

若$\hat{\boldsymbol{\pi}}\xrightarrow{p}\boldsymbol{\pi}_0$且$\mathbf{W}_N\xrightarrow{p}\mathbf{W}_0$，其中，$\mathbf{W}_0$表示有限半正定矩阵，则$Q_N(\hat{\boldsymbol{\theta}})\xrightarrow{p}Q_0(\boldsymbol{\theta})=(\boldsymbol{\pi}_0-\mathbf{g}(\boldsymbol{\theta}))'\mathbf{W}_0(\boldsymbol{\pi}_0-\mathbf{g}(\boldsymbol{\theta}))$。由此可得，当$\mathrm{Rank}[\mathbf{W}_0\times\partial\mathbf{g}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}']=q$，则$\theta_0$是局部可识别的，而一致性本质上要求$\boldsymbol{\pi}_0=\mathbf{g}(\boldsymbol{\theta}_0)$。

对最小距离估计量来说，$\sqrt{N}(\hat{\boldsymbol{\theta}}_{\mathrm{MD}}-\boldsymbol{\theta}_0)\xrightarrow{d}\mathcal{N}[\mathbf{0},\ \mathrm{V}[\hat{\boldsymbol{\theta}}_{\mathrm{MD}}]]$，其中：

$$\mathrm{V}[\hat{\boldsymbol{\theta}}_{\mathrm{MD}}]=(\mathbf{G}_0'\mathbf{W}_0\mathbf{G}_0)^{-1}(\mathbf{G}_0'\mathbf{W}_0\mathrm{V}[\hat{\boldsymbol{\pi}}]\mathbf{W}_0\mathbf{G}_0)(\mathbf{G}_0'\mathbf{W}_0\mathbf{G}_0)^{-1} \tag{6.67}$$

$\mathbf{G}_0=\partial g(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'|_{\theta_0}$，同时假定简化式参数$\hat{\boldsymbol{\pi}}$具有极限分布$\sqrt{N}(\hat{\boldsymbol{\pi}}-\boldsymbol{\pi}_0)\xrightarrow{d}\mathcal{N}[\mathbf{0},\mathrm{V}[\hat{\boldsymbol{\pi}}]]$。由于较小的$\mathrm{V}[\hat{\boldsymbol{\pi}}]$会使式(6.67)中的$\mathrm{V}[\hat{\boldsymbol{\theta}}_{\mathrm{MD}}]$较小，所以更有效的简化式估计量会产生更有效的最小距离估计量。

为了得到结果(6.67)，以下述对最小距离估计量的一阶条件重新标度来开始：

$$G_N(\hat{\boldsymbol{\theta}})'\mathbf{W}_N\sqrt{N}(\hat{\boldsymbol{\pi}}-\mathbf{g}(\hat{\boldsymbol{\theta}}))=\mathbf{0} \tag{6.68}$$

其中，$G_N(\boldsymbol{\theta})=\partial\mathbf{g}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'$。在$\boldsymbol{\theta}_0$附近进行精确一阶泰勒级数展开，即：

$$\sqrt{N}\mathbf{h}(\hat{\boldsymbol{\pi}}-\mathbf{g}(\hat{\boldsymbol{\theta}}))=\sqrt{N}(\hat{\boldsymbol{\pi}}-\boldsymbol{\pi}_0)-G_N(\boldsymbol{\theta}^+)\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) \quad (6.69)$$

其中，$\boldsymbol{\theta}^+$ 位于 $\hat{\boldsymbol{\theta}}$ 与 $\boldsymbol{\theta}_0$ 之间，同时使用了 $\mathbf{g}(\boldsymbol{\theta}_0)=\boldsymbol{\pi}_0$。将式(6.69)代入式(6.68)，并解出 $\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$，得到：

$$\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)=[G_N(\hat{\boldsymbol{\theta}})'\mathbf{W}_N G_N(\boldsymbol{\theta}^+)]^{-1}G_N(\hat{\boldsymbol{\theta}})'\mathbf{W}_N\sqrt{N}(\hat{\boldsymbol{\pi}}-\boldsymbol{\pi}_0) \quad (6.70)$$

这就直接得出式(6.67)。

给定简化式估计量 $\hat{\boldsymbol{\pi}}$，最有效的最小距离估计量运用式(6.66)的加权矩阵 $\mathbf{W}_N=\hat{\mathrm{V}}[\hat{\boldsymbol{\pi}}]^{-1}$。此估计量称为最优最小距离估计量[**optimal/MD (OMD) estimator**]，有时遵从弗格森(Ferguson, 1958)的说法，称之为最小卡方估计量(**minimum chi-square estimator**)。

一种普遍的可供选择的特殊情况是等价加权最小距离估计量[**equally weighted minimum distance (EWMD) estimator**]，它设 $\mathbf{W}_N=\mathbf{I}$。该估计量的有效性比最优最小距离估计量稍差一些，但它并不具有有限样本偏倚问题，这点类似于 6.3.5 节曾经讨论的当运用最优加权矩阵时产生的那些问题。

等价加权最小距离估计量能通过 $\hat{\pi}_j$ 对 $g_j(\hat{\boldsymbol{\theta}})(j=1,\cdots,r)$ 的 NLS 回归而直接获得，因此，极小化 $(\hat{\boldsymbol{\pi}}-\mathbf{g}(\hat{\boldsymbol{\theta}}))'(\hat{\boldsymbol{\pi}}-\mathbf{g}(\hat{\boldsymbol{\theta}}))$ 时产生的一阶条件，与具有 $\mathbf{W}_N=\mathbf{I}$ 的式(6.68)中的一阶条件相同。

对最优最小距离的目标函数求极大值，就得出卡方分布。特别地，有：

$$(\hat{\boldsymbol{\pi}}-\mathbf{g}(\hat{\boldsymbol{\theta}}_{\mathrm{OMD}}))'\hat{\mathrm{V}}[\hat{\boldsymbol{\pi}}]^{-1}(\hat{\boldsymbol{\pi}}-\mathbf{g}(\hat{\boldsymbol{\theta}}_{\mathrm{OMD}})) \quad (6.71)$$

上式在 H_0：$\mathbf{g}(\boldsymbol{\theta}_0)=\boldsymbol{\pi}_0$ 下渐近服从 $\chi^2(r-q)$。这提供类似于 6.3.8 节中 OIR 检验的一种模型设定检验。

最小距离估计量在性质上类似于广义矩方法估计量。广义矩方法框架是一种被广泛使用的标准框架。最小距离估计经常用于协方差结构的面板研究中，这是因为，$\hat{\boldsymbol{\pi}}$ 包含很容易估计的样本矩(方差与协方差)，而这些样本矩用于得出 $\hat{\boldsymbol{\theta}}$。

6.8 经验似然法

矩方法与广义矩方法并不要求对条件密度的完全设定。可是，估计可以建立在形式为 $\mathrm{E}[\mathbf{h}(y,\mathbf{x},\boldsymbol{\theta})]=\mathbf{0}$ 的矩条件基础上。归功于欧文(Owen, 1988)的经验似然方法，则是建立在同样的矩条件基础上的一种可供选择的估计方法。

尽管经验似然估计量在渐近形式上等价于广义矩方法估计量，但其引人注目的地方是，它具有不同的有限样本性质，并且在一些例子中超过了广义矩方法估计量。

6.8.1 *总体均值经验似然估计*

我们以纯量 iid 随机变量 y 的情况开始讨论，其中，y 具有密度 $f(y)$ 以及样本似然函数 $\Pi_i f(y_i)$。这里所考虑的复杂情况是没有设定密度 $f(y)$，因而不可以运用通常的极大似然方法。

完全非参数方法企图在 y 的第 i 个样本值处估计密度 $f(y)$。设 $\pi_i=f(y_i)$ 表示 y 的第 i 个观测值取实现值 y_i 的概率。其目标是对所谓的经验似然函数 $\Pi_i\pi_i$ 求极大值，或者等价地，对对数经验似然函数 $N^{-1}\sum_i \ln\pi_i$ 求极大值，这是对 π_i 没有施加结构的多项式模型。该对数似然是无界的，除非对 π_i 的取值范围加上一个约束。一个常用的正规化是 $\sum_i\pi_i=1$。在完全非参数的情况下，正如我们现在所阐述的，得出累积分布函数的标准估计。

经验似然估计量极大化 $\boldsymbol{\pi}$ 与 η 的拉格朗日算子：

$$\mathcal{L}_{\mathrm{EL}}(\boldsymbol{\pi},\eta)=\frac{1}{N}\sum_{i=1}^{N}\ln\pi_i-\eta\left(\sum_{i=1}^{N}\pi_i-1\right) \tag{6.72}$$

其中，$\boldsymbol{\pi}=[\pi_1,\cdots,\pi_N]'$，而 η 表示拉格朗日乘子。尽管数据 y_i 并没有明确出现在式(6.72)中，但 y_i 却以隐性方式出现并成为 $\pi_i=f(y_i)$。将 $\pi_i(i=1,\cdots,N)$ 与 η 的导数设定为零，然后求解它们，得到 $\hat{\pi}_i=1/N$ 与 $\eta=1$。因此，估计密度函数 $\hat{f}(y)$ 在 y_i 的每一个实现值上具有质量 $1/N,i=1,\cdots,N$。所得到的分布函数是 $\hat{F}(y)=N^{-1}\sum_{i=1}^{N}\mathbf{1}(y\leqslant y_i)$，其中，当事件 A 发生时，有 $\mathbf{1}(A)=1$，否则 $\mathbf{1}(A)=0$。$\hat{F}(y)$ 恰好是通常的经验分布函数。

现在引入一些参数。举一个简单例子，假如我们引进矩约束 $\mathrm{E}[y-\mu]=0$，其中，μ 表示未知的总体均值。在经验似然背景下，这个总体矩可用样本矩代替，其中，样本矩是通过概率 π_i 来对样本值进行加权。因此，我们引入约束 $\sum_i\pi_i(y_i-\mu)=0$。经验极大似然估计量的拉格朗日算子是：

$$\mathcal{L}_{\mathrm{EL}}(\boldsymbol{\pi},\eta,\lambda,\mu)=\frac{1}{N}\sum_{i=1}^{N}\ln\pi_i-\eta\left(\sum_{i=1}^{N}\pi_i-1\right)-\lambda\sum_{i=1}^{N}\pi_i(y_i-\mu) \tag{6.73}$$

其中，η 与 λ 均表示拉格朗日乘子。

我们从对 $\pi_i(i=1,\cdots,N)$、η 和 λ 而不是 μ 求拉格朗日算子的导数开始。令这些导数为 0，则得到作为 μ 的函数的一些方程。然后解方程，得出其解 $\pi_i=\pi_i(\mu)$，进而获得对 μ 求极大值的经验似然 $N^{-1}\sum_i\ln\pi_i(\mu)$。这种求解方法所得到的非线性方程，需要用数值方法加以求解。

对这个特殊问题来说，解出 μ 的一种比较容易的方法是，注意到，$\mathcal{L}(\boldsymbol{\pi},\eta,\lambda,\mu)$ 的极大值必须小于或等于 $N^{-1}\sum_i\ln N^{-1}$，这是因为，它是一个没有最终约束的极大值。不过，若 $\pi_i=1/N$ 且 $\hat{\mu}=N^{-1}\sum_i y_i=\bar{y}$，则 $\mathcal{L}(\boldsymbol{\pi},\eta,\lambda,\mu)$ 等于 $N^{-1}\sum_i\ln N^{-1}$。因此，总体均值的经验极大似然估计量就是样本均值。

6.8.2 回归参数经验似然估计

现在，考察随 i 而变化的 iid 的回归数据。对此模型施加的唯一结构是 r 个矩条件：

$$\mathrm{E}[\mathbf{h}(\mathbf{w}_i,\boldsymbol{\theta})]=\mathbf{0} \tag{6.74}$$

其中，$\mathrm{h}(\cdot)$ 与 w_i 都已由 6.3.1 节定义。例如，对于 OLS 估计来说，$\mathbf{h}(\mathbf{w},\boldsymbol{\theta})=\mathbf{x}(y-\mathbf{x}'\boldsymbol{\theta})$；而对于 NLS 估计来说，$\mathbf{h}(y,\mathbf{x},\boldsymbol{\theta})=(\partial g/\partial\boldsymbol{\theta})(y-g(\mathbf{x},\boldsymbol{\theta}))$。

经验似然方法就是对经验似然函数 $N^{-1}\sum_i \ln\pi_i$ 求极大值，其约束为 $\sum_i \pi_i=1$ [参见(6.72)]以及建立在总体矩条件(6.74)基础上的另外样本约束，即：

$$\sum_{i=1}^{N}\pi_i\mathbf{h}(\mathbf{w}_i,\boldsymbol{\theta})=\mathbf{0} \tag{6.75}$$

因此，我们对 $\boldsymbol{\pi}$、η、$\boldsymbol{\lambda}$ 以及 $\boldsymbol{\theta}$ 求极大值：

$$\mathcal{L}_{\mathrm{EL}}(\boldsymbol{\pi},\eta,\boldsymbol{\lambda},\boldsymbol{\theta})=\frac{1}{N}\sum_{i=1}^{N}\ln\pi_i-\eta\Big(\sum_{i=1}^{N}\pi_i-1\Big)-\boldsymbol{\lambda}'\sum_{i=1}^{N}\pi_i\mathbf{h}(\mathbf{w}_i,\boldsymbol{\theta}) \tag{6.76}$$

其中，拉格朗日算子是与 $\mathbf{h}(\cdot)$ 维数相同的纯量 η 和列向量 $\boldsymbol{\lambda}$。

首先，关注 N 个数 $\pi_1,\cdots,\pi_N$。对 $\mathcal{L}(\boldsymbol{\pi},\eta,\boldsymbol{\lambda},\boldsymbol{\theta})$ 求关于 π_i 的微分，得到 $1/(N\pi_i)-\eta-\boldsymbol{\lambda}'\mathbf{h}_i=0$。于是，用 π_i 去乘并对 i 求和，再利用 $\sum_i\pi_i\mathbf{h}_i=\mathbf{0}$，从而得到 $\eta=1$。由此可得：

$$\pi_i(\boldsymbol{\theta},\boldsymbol{\lambda})=\frac{1}{N(1+\boldsymbol{\lambda}'\mathbf{h}(\mathbf{w}_i,\boldsymbol{\theta}))} \tag{6.77}$$

该问题现在简化成为求关于 $(r+q)$ 个变量 $\boldsymbol{\lambda}$ 与 $\boldsymbol{\theta}$ 的极大值问题，而且，其拉格朗日算子与 r 个矩条件(6.74)以及 q 个参数 $\boldsymbol{\theta}$ 有关。

甚至对恰好识别模型来说，需要用数值方法加以求解。人们对函数 $N^{-1}\sum_i\times\ln[1/N(1+\boldsymbol{\lambda}'\mathbf{h}_i(\mathbf{w}_i,\boldsymbol{\theta}))]$ 求关于 $\boldsymbol{\theta}$ 与 $\boldsymbol{\lambda}$ 的极大值。

或者，首先关注 $\boldsymbol{\lambda}$。对 $\mathcal{L}(\boldsymbol{\pi}(\boldsymbol{\theta},\boldsymbol{\lambda}),\eta,\boldsymbol{\lambda})$ 求关于 $\boldsymbol{\lambda}$ 的微分，得到 $\sum_i\pi_i\mathbf{h}_i=\mathbf{0}$。把 $\boldsymbol{\lambda}(\boldsymbol{\theta})$ 定义为 $\dim(\boldsymbol{\lambda})$ 方程组

$$\sum_{i=1}^{N}\frac{1}{N(1+\boldsymbol{\lambda}'\mathbf{h}(\mathbf{w}_i,\boldsymbol{\theta}))}\mathbf{h}(\mathbf{w}_i,\boldsymbol{\theta})=\mathbf{0}$$

的隐性解。对其求解时要使用数值方法，进而得出 $\boldsymbol{\lambda}(\boldsymbol{\theta})$。于是，式(6.77)变为：

$$\pi_i(\boldsymbol{\theta})=\frac{1}{N(1+\boldsymbol{\lambda}(\boldsymbol{\theta})'\mathbf{h}(\mathbf{w}_i,\boldsymbol{\theta}))} \tag{6.78}$$

通过把式(6.78)代入经验似然函数 $N^{-1}\sum_i\ln\pi_i$ 中，经验对数似然函数在 $\boldsymbol{\theta}$ 处的计算值是：

$$\mathcal{L}_{\mathrm{EL}}(\boldsymbol{\theta})=-\frac{1}{N}\sum_{i=1}^{N}\ln[N(1+\boldsymbol{\lambda}(\boldsymbol{\theta})'\mathbf{h}(\mathbf{w}_i,\boldsymbol{\theta}))]$$

求这个函数关于 $\boldsymbol{\theta}$ 的极大值，即为经验极大似然估计量[**maximum empirical likelihood (MEL) estimator**] $\hat{\boldsymbol{\theta}}_{\mathrm{MEL}}$。

秦和劳利斯(Qin and Lawless, 1994)已经证明：

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{\mathrm{MEL}}-\boldsymbol{\theta}_0)\xrightarrow{d}\mathcal{N}[\mathbf{0},\mathbf{A}(\boldsymbol{\theta}_0)^{-1}\mathbf{B}(\boldsymbol{\theta}_0)\mathbf{A}(\boldsymbol{\theta}_0)'^{-1}]$$

其中，$\mathbf{A}(\boldsymbol{\theta}_0)=\mathrm{plim}\,\mathrm{E}[\partial\mathbf{h}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'|_{\theta_0}]$，而 $\mathbf{B}(\boldsymbol{\theta}_0)=\mathrm{plim}\,\mathrm{E}[\mathbf{h}(\boldsymbol{\theta})\mathbf{h}(\boldsymbol{\theta})'|_{\theta_0}]$。这与矩方法[参见式(6.13)]的分布相同。不过，在有限样本中，$\hat{\boldsymbol{\theta}}_{\mathrm{MEL}}$ 与 $\hat{\boldsymbol{\theta}}_{\mathrm{GMM}}$ 却有所不同，其推断建立在样本估计值

$$\hat{\mathbf{A}} = \sum\nolimits_{i=1}^{N} \hat{\pi}_i \frac{\partial \mathbf{h}_i'}{\partial \boldsymbol{\theta}} \bigg|_{\hat{\theta}}$$

$$\hat{\mathbf{B}} = \sum\nolimits_{i=1}^{N} \hat{\pi}_i \mathbf{h}_i(\hat{\boldsymbol{\theta}})\mathbf{h}_i(\hat{\boldsymbol{\theta}})'$$

的基础上，通过估计概率 $\hat{\pi}_i$ 而不是比例 $1/N$ 来进行加权。

英伯斯(Imbens, 2002)曾经提供将经验似然法与广义矩方法进行比较的经验似然法的一个最新综述。一些变形包括，通过 $N^{-1}\sum_i \pi_i \ln \pi_i$ 代替式(6.76)的 $N^{-1}\sum_i \ln \pi_i$。经验似然法在计算上更为繁琐；有关讨论，参见英伯斯(Imbens, 2002)。其优点是，渐近理论研究表明，与广义矩方法估计量的有限样本近似相比，经验似然估计量分布的有限样本近似表现得更好。

6.9 线性方程组

上述估计理论涵盖了大多数应用研究所使用的单方程估计方法。现在，我们考察几个方程的联合估计。本节阐述具有可加误差的关于参数为线性的一些方程，而下一节则给出对非线性方程组的推广。

联合估计的主要优点是提高有效性，这是因为对给定个体来说，并入了不可观测的交叉方程方面的相关性。再者，若交叉方程系数存在约束，就必须进行联合估计。对外生回归元方程组来说，估计是对单方程 OLS 与 GLS 估计的稍微推广；而对内生回归元来说，估计则是改进的单方程工具变量方法。

对许多个体而言，一个重要例子是在某个时点上的那些可观测的几种商品的需求方程组。对看似不相关回归来说，所有回归元都是外生的；而对联立方程模型来说，一些回归元是内生的。

第二个重要例子是面板数据，其中，对许多个体而言，在几个时点上的单个方程都是可观测的，并且把每一个时期处理成为单独方程。通过把面板数据模型看成是方程组的例子，当某些回归元是内生的时候，改进有效性、获得面板标准误差以及推导工具就是可行的。

许多经济计量学教科书都对线性方程组内容进行长篇大论。这里的阐述则非常简洁。此处主要针对非线性方程组的推广(参见 6.10 节)以及面板数据的应用(参见第 21 章～第 23 章)。

6.9.1 *线性方程组*

单方程线性模型由 $y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$ 给出，其中，y_i 与 u_i 均表示纯量，而 $\mathbf{x}_i$ 与 $\boldsymbol{\beta}$ 均表示列向量。具有 G 个因变量的多方程线性模型(**multiple-equation linear model**)或多元变量线性模型(**multivariate linear model**)由

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i, \quad i = 1, \cdots, N \tag{6.79}$$

给出，其中，$\mathbf{y}_i$ 与 $\mathbf{u}_i$ 均表示 $G\times 1$ 维向量，而 $\mathbf{X}_i$ 表示 $G\times K$ 阶矩阵，$\boldsymbol{\beta}$表示 $K\times 1$ 维列向量。

本节始终做出误差向量 $\mathbf{u}_i$ 对于不同 i 都是独立的横截面假设，因此 $\mathrm{E}[\mathbf{u}_i\mathbf{u}_j'] =$

$\mathbf{0}$，对于 $i \neq j$。不过，对于给定 i 来说，$\mathbf{u}_i$ 的成分可能是相关的，而且其方差与协方差随 i 而变化，就第 i 个个体而言，得出条件误差矩阵：

$$\boldsymbol{\Omega}_i = \mathrm{E}[\mathbf{u}_i \mathbf{u}_i' | \mathbf{X}_i] \tag{6.80}$$

存在多种方式建立多方程模型。在一种极端情况下，看似不相关方程模型把 G 个方程组合起来，诸如对各种不同消费者而言的商品需求，其中，参数会随不同方程而变化，而回归元对于不同方程来说可能变化也可能不变化。在另一种极端情况下，线性面板数据则把相同方程的 G 个时期数据组合起来，其参数在不同时期为常值，并且回归元在不同时期可能变化也可能不变化。这两种情况将在 6.9.3 节与 6.9.4 节加以阐述。

若对 N 个个体叠放式(6.79)，则得到：

$$\begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_N \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_N \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_N \end{bmatrix} \tag{6.81}$$

或

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{6.82}$$

其中，$\mathbf{y}$ 与 $\mathbf{u}$ 均表示 $NG \times 1$ 维向量，$\mathbf{X}$ 表示 $NG \times K$ 阶矩阵。

下面给出的结果如同单方程情况一样，以同样方式处理叠放模型(6.82)而获得。因此，OLS 估计量是 $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$，而在具有工具矩阵 Z 的恰好识别情况下，工具变量估计量是 $\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$。其唯一变化是，对角误差矩阵的横截面假设由分块对角误差矩阵假设来代替。在计算系统估计量的估计方差矩阵和构建可行 GLS 估计量与有效广义矩方法估计量时，都需要考虑这种对角性。

6.9.2 *系统 OLS 与 FGLS 估计*

对方程组(6.82)进行 OLS 估计，得到系统普通最小二乘法估计量(**systems OLS estimator**)$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$。利用式(6.81)，可立刻得出：

$$\hat{\boldsymbol{\beta}}_{\mathrm{SOLS}} = \left[\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{X}_i\right]^{-1} \sum_{i=1}^{N} \mathbf{X}_i'\mathbf{y}_i \tag{6.83}$$

该估计量服从渐近正态分布，若假定数据对不同 i 是独立的，就可利用通常稳健三明治结果，从而有：

$$\hat{\mathrm{V}}[\hat{\boldsymbol{\beta}}_{\mathrm{SOLS}}] = \left[\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{X}_i\right]^{-1} \sum_{i=1}^{N} \mathbf{X}_i'\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'\mathbf{X}_i \left[\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{X}_i\right]^{-1} \tag{6.84}$$

其中，$\hat{\mathbf{u}}_i = \mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}$。此方差矩阵估计值允许误差的条件方差与协方差随不同个体的变化而不同。

给定个体且已知误差向量成分的相关性，则更有效的估计可通过 GLS 或 FGLS 来获得。若观测值对不同 i 是独立的，系统 GLS 估计量(**systems GLS estimator**)就可应用到变换方程组：

$$\mathbf{\Omega}_i^{-1/2}\mathbf{y}_i=\mathbf{\Omega}_i^{-1/2}\mathbf{X}_i\boldsymbol{\beta}+\mathbf{\Omega}_i^{-1}\mathbf{u}_i \tag{6.85}$$

其中，$\mathbf{\Omega}_i$ 表示由式(6.80)定义的误差方差矩阵。变换误差 $\mathbf{\Omega}_i^{-1/2}\mathbf{u}_i$ 具有零均值，其方差为：

$$\begin{aligned}\mathrm{E}[(\mathbf{\Omega}_i^{-1/2}\mathbf{u}_i)'(\mathbf{\Omega}_i^{-1/2}\mathbf{u}_i)\mid\mathbf{X}_i]&=\mathbf{\Omega}_i^{-1/2}\mathrm{E}[\mathbf{u}_i'\mathbf{u}_i\mid\mathbf{X}_i]\mathbf{\Omega}_i^{-1/2}\\&=\mathbf{\Omega}_i^{-1/2}\mathbf{\Omega}_i\mathbf{\Omega}_i^{-1/2}\\&=\mathbf{I}_G\end{aligned}$$

因此，变换方程组的误差为同方差的，且对 G 个方程来说，是不相关的，从而 OLS 是有效的。

为了得到此估计量，需要对 $\mathbf{\Omega}_i$ 模型加以设定，比如说 $\mathbf{\Omega}_i=\mathbf{\Omega}_i(\gamma)$。然后，对变换方程组执行系统 OLS 估计，其中，$\mathbf{\Omega}_i$ 用 $\mathbf{\Omega}_i(\hat{\gamma})$代替，而 $\hat{\gamma}$ 表示 γ 的一致估计值。这就得到系统可行广义最小二乘法估计量[**system feasible GLS (SFGLS) estimator**]：

$$\hat{\boldsymbol{\beta}}_{\mathrm{SFGLS}}=\left[\sum_{i=1}^{N}\mathbf{X}_i'\hat{\mathbf{\Omega}}_i^{-1}\mathbf{X}_i\right]^{-1}\sum_{i=1}^{N}\mathbf{X}_i'\hat{\mathbf{\Omega}}_i^{-1}\mathbf{y}_i \tag{6.86}$$

该估计量服从渐近正态分布，同时为了防止对 $\mathbf{\Omega}_i(\gamma)$的可能错误设定，可使用方差矩阵的稳健三明治估计值：

$$\hat{\mathrm{V}}[\hat{\boldsymbol{\beta}}_{\mathrm{SFGLS}}]=\left[\sum_{i=1}^{N}\mathbf{X}_i'\hat{\mathbf{\Omega}}_i^{-1}\mathbf{X}_i\right]^{-1}\sum_{i=1}^{N}\mathbf{X}_i'\hat{\mathbf{\Omega}}_i^{-1}\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'\hat{\mathbf{\Omega}}_i^{-1}\mathbf{X}_i\left[\sum_{i=1}^{N}\mathbf{X}_i'\hat{\mathbf{\Omega}}_i^{-1}\mathbf{X}_i\right]^{-1} \tag{6.87}$$

其中，$\mathbf{\Omega}_i=\mathbf{\Omega}_i(\hat{\gamma})$。

对 $\mathbf{\Omega}_i$ 的最普遍设定是，对不同 i 来说，假定 $\mathbf{\Omega}_i$ 是不变的。那么，就有限的 G 且 $N\to\infty$而言，$\mathbf{\Omega}_i=\mathbf{\Omega}$ 是一个 $G\times G$ 阶矩阵，$\mathbf{\Omega}$ 可通过

$$\hat{\mathbf{\Omega}}=\frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i' \tag{6.88}$$

得到一致估计，其中 $\hat{\mathbf{u}}_i=\mathbf{y}_i-\mathbf{X}_i\hat{\boldsymbol{\beta}}_{\mathrm{SOLS}}$。于是，式(6.86)的系统可行广义最小二乘法估计量就是用 $\hat{\mathbf{\Omega}}$ 代替 $\hat{\mathbf{\Omega}}_i$，经过一些代数运算之后，还可将此估计量写成：

$$\hat{\boldsymbol{\beta}}_{\mathrm{SFGLS}}=[\mathbf{X}'(\hat{\mathbf{\Omega}}^{-1}\otimes\mathbf{I}_N)\mathbf{X}]^{-1}\mathbf{X}'(\hat{\mathbf{\Omega}}^{-1}\otimes\mathbf{I}_N)\mathbf{y}' \tag{6.89}$$

其中，$\otimes$表示克罗内克积(Kronecker product)。例如，对不同 i 的异方差性来说，要排除其假设：$\mathbf{\Omega}_i=\mathbf{\Omega}$。这是一个很强的假设，而且在许多应用中，一种最好的方式是利用式(6.87)计算稳健标准误差，即使 $\mathbf{\Omega}_i$ 随不同 i 而变化，仍能得到正确的标准误差。

6.9.3 *看似不相关回归*

看似不相关回归模型[**seemingly unrelated regression (SUR) model**]设定如下，对 N 个个体的第 i 个而言，G 个方程的第 g 个是由下式给出：

$$y_{ig}=\mathbf{x}_{ig}'\boldsymbol{\beta}_g+u_{ig},\quad g=1,\cdots,G,\ i=1,\cdots,N \tag{6.90}$$

其中，$\mathbf{x}_{ig}$表示回归元，假定 $\mathbf{x}_{ig}$是外生的，$\boldsymbol{\beta}_g$ 表示 $K_g\times1$ 维参数向量。例如，有 N

个个体需求 G 种商品数据，y_{ig} 可以是第 i 个个体对商品 g 的开支，或者是对商品 g 的预算值。尽管 $N\to\infty$，但假定合计总数 G 是固定的且适当小。注意到，我们使用下标次序 y_{ig} 作为结果，就很容易对具有变量 y_{it} 的面板数据加以变换（参见 6.9.4 节）。其他一些学者则使用相反次序 y_{gi}。

看似不相关回归模型是由泽尔纳（Zellner，1962）提出的。看似不相关回归这一术语容易使人产生误解，因为如果不同方程中的误差 u_{ig} 是相关的，那么一些方程显然是相关的。对看似不相关回归模型来说，y_{ig} 与 y_{ih} 之间的关系是间接的；这会通过相关关系而转递到不同方程的误差关系上。

估计是把不同方程的观测值与每个个体的观测值结合起来。从微观经济计量学应用的角度来看，假定对不同 i 具有独立性，一种最简便的方式是，首先对给定个体叠放所有方程。对第 i 个个体的所有 G 个方程加以叠放，得出：

$$\begin{bmatrix} y_{i1} \\ \vdots \\ y_{iG} \end{bmatrix} = \begin{bmatrix} \mathbf{x}'_{i1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{x}'_{iG} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_G \end{bmatrix} + \begin{bmatrix} u_{i1} \\ \vdots \\ u_{iG} \end{bmatrix} \tag{6.91}$$

它具有式(6.79)的 $\mathbf{y}_i=\mathbf{X}_i\boldsymbol{\beta}+\mathbf{u}_i$ 形式，其中，$\mathbf{y}_i$ 与 $\mathbf{u}_i$ 均表示 $G\times 1$ 维向量，其第 g 个元素分别为 y_{ig} 与 u_{ig}，$\mathbf{X}_i$ 表示 $G\times K$ 阶矩阵，其第 g 行为 $[\mathbf{0}\cdots\mathbf{x}'_{ig}\cdots\mathbf{0}]$，而 $\boldsymbol{\beta}=[\boldsymbol{\beta}'_1\cdots\boldsymbol{\beta}'_i]$ 表示 $K\times 1$ 维向量，其中，$K=K_1+\cdots+K_G$。不过，一些作者对给定方程进行叠放，得出相同估计量，但具有不同的代数表达式。

已知 $\mathbf{X}_i$ 与 $\mathbf{y}_i$ 的定义，容易证明式(6.83)中的 $\hat{\boldsymbol{\beta}}_{\mathrm{SOLS}}$ 是：

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \vdots \\ \hat{\boldsymbol{\beta}}_G \end{bmatrix} = \begin{bmatrix} \left[\sum_{i=1}^{N}\mathbf{x}_{i1}\mathbf{x}'_{i1}\right]^{-1}\sum_{i=1}^{N}\mathbf{x}_{i1}y_{i1} \\ \vdots \\ \left[\sum_{i=1}^{N}\mathbf{x}_{iG}\mathbf{x}'_{iG}\right]^{-1}\sum_{i=1}^{N}\mathbf{x}_{iG}y_{iG} \end{bmatrix}$$

因此，系统 OLS 与各自逐一方程 OLS（**equation-by-equation OLS**）是一样的。正如先前人们所料，若不同方程之间的唯一联系是误差，同时误差可被处理成不相关的，则联合估计就简化成为单方程估计。

一个较好的估计量是由式(6.86)所定义的可行 GLS 估计量，它利用式(6.88)中的 $\hat{\boldsymbol{\Omega}}$ 和基于式(6.87)中渐近方差的统计推断。一般来讲，此估计量比系统 OLS 更为有效，尽管可以证明，若误差在不同方程之间是不相关的，或者相同回归元恰好出现在每一个方程中，则会简化成 OLS。

看似不相关回归模型可利用交叉方程参数约束（**cross-equation parameter restrictions**）。例如，对称性约束可能蕴含着，第一个方程中的第二个回归元系数等于第二个方程中的第一个回归元的系数。如果这类约束是等式约束，那么人们很容易通过式(6.79)给出的 $\mathbf{X}_i$ 与 $\boldsymbol{\beta}$ 重新适当定义来估计模型。例如，如果存在两个方程且约束是 $\boldsymbol{\beta}_2=-\boldsymbol{\beta}_1$，那么定义 $\mathbf{X}_i=[\mathbf{x}_{i1}\ \ -\mathbf{x}_{i2}]'$ 与 $\boldsymbol{\beta}=\boldsymbol{\beta}_1$。或者，利用对其参数为线性约束的单方程 OLS 与具有 GLS 的方程组进行推广来加以估计。

此外，方程组可能出现的情况是，误差向量 u_i 的方差矩阵是奇异的，这是由加总约束（**adding-up constraints**）引起的。例如，假定 y_{ig} 表示第 i 个预算值，从而模型

$y_{ig}=\alpha_g+\mathbf{z}_g'\boldsymbol{\beta}_g+u_{ig}$，其中，相同回归元出现在每一个方程中。那么，由于预算值之和为 1，所以 $\sum_g y_{ig}=1$，这就要求 $\sum_g\alpha_g=1$、$\sum_g\beta_g=\mathbf{0}$ 以及 $\sum_g u_{ig}=0$。而最后的约束意味着 $\boldsymbol{\Omega}_i$ 是奇异的，从而是不可逆的。人们能去掉一个方程，比如说最后一个，然后通过对剩下的 $G-1$ 方程用系统估计法估计模型。于是，第 G 个方程的参数估计可利用加总约束来获得。例如，$\hat{\alpha}_G=1-(\hat{\alpha}_1+\cdots+\hat{\alpha}_{G-1})$。在此背景下，对参数利用等式约束也是可行的。文献中存在一些方法：所求估计值在去掉方程后是不变的。例如，参见波恩特和萨文(Berndt and Savin，1975)。

6.9.4 面板数据

系统 GLS 方法的另一个重要应用是面板数据，其中，对 N 个个体来说，纯量因变量在 T 个时期的每一个中都是可观测的。可将面板数据看成一个方程组，即 N 个个体的 T 个方程或者 T 个时期的 N 个方程。在微观经济计量学中，我们假定短面板具有很小的 T 且 $N\rightarrow\infty$，故把它设置成纯量因变量 y_{it} 是很自然的，其中，前面所讨论的第 g 个方程现在被解释为在第 t 个时期且 $G=T$。

一个简单面板数据模型(**simple panel data model**)是：

$$y_{it}=\mathbf{x}_{it}'\boldsymbol{\beta}+u_{it},\quad t=1,\cdots,T,\ i=1,\cdots,N \tag{6.92}$$

它是式(6.90)的一种特殊形式，其中，$\boldsymbol{\beta}$ 为常值。于是，式(6.79)的回归元矩阵变成 $\mathbf{X}_i=[\mathbf{x}_{i1}\cdots\mathbf{x}_{iT}]'$。经过一些代数运算后，式(6.83)定义的系统 OLS 估计量能重新写成：

$$\hat{\boldsymbol{\beta}}_{\text{POLS}}=\left[\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{x}_{it}\mathbf{x}_{it}'\right]^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{x}_{it}y_{it} \tag{6.93}$$

该估计量称为混合 OLS 估计量(**pooled OLS estimator**)，这是因为它把横截面数据与时间序列数据混合或组合在一起。

混合估计量可直接通过 y_{it} 对 $\mathbf{x}_{it}$ 的 OLS 估计来获得。不过，若对于给定的 i，u_{it} 对不同的 t 是相关的，则被假定为既对不同 i 又对不同 t 具有误差独立性的默认 OLS 标准误差是无效的，且具有很大的向下偏倚。但是，统计推断应建立在由式(6.84)给出的协方差矩阵的稳健形式的基础上。这将在 21.2.3 节详细阐述。在实际应用中，可估计比包括特定个体效应的式(6.92)更为复杂的模型(参见 21.2 节)。

6.9.5 系统工具变量估计

6.4 节已经阐述对具有内生回归元的单个线性方程估计。当 $\mathrm{E}[\mathbf{u}_i|\mathbf{X}_i]\neq\mathbf{0}$ 时，将这种方法推广到多元线性模型(6.79)上。为了获得一致且有效的估计，布伦迪和乔根森(Brundy and Jorgenson，1971)曾考察用于方程组的工具变量估计。

我们假定存在 $G\times r$ 阶工具矩阵 $\mathbf{Z}_i$，它满足 $\mathrm{E}[\mathbf{u}_i|\mathbf{Z}_i]=\mathbf{0}$，因此有：

$$\mathrm{E}[\mathbf{Z}_i'(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})]=\mathbf{0} \tag{6.94}$$

若利用单方程工具变量方法，这些工具能用于获得一致参数估计，但联合方程估计能改进有效性。系统广义矩方法估计量(**systems GMM estimator**)极小化下式：

$$Q_N(\beta)=\left[\sum_{i=1}^{N}\mathbf{Z}_i'(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})\right]'\mathbf{W}_N\left[\sum_{i=1}^{N}\mathbf{Z}_i'(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})\right] \tag{6.95}$$

其中，$\mathbf{W}_N$ 表示 $r\times r$ 阶加权矩阵。经过一些代数运算，得到：

$$\hat{\boldsymbol{\beta}}_{SGMM}=[\mathbf{X}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\mathbf{X}]^{-1}[\mathbf{X}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\mathbf{y}]^{-1} \tag{6.96}$$

其中，$\mathbf{X}$ 表示通过叠放 $\mathbf{X}_1,\cdots,\mathbf{X}_N$ 获得的 $NG\times K$ 阶矩阵[参见式(6.81)]，而 $\mathbf{Z}$ 表示通过类似方式叠放获得的 $NG\times r$ 阶矩阵。系统广义矩方法估计量确实与式(6.37)具有相同形式，而且其渐近方差矩阵是由式(6.39)给出的形式。由此可得，在方程组情况下，同时假定对不同 i 具有独立性，其方差矩阵的稳健估计值是：

$$\hat{\mathrm{V}}[\hat{\boldsymbol{\beta}}_{SGMM}]=N[\mathbf{X}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\mathbf{X}]^{-1}[\mathbf{X}'\mathbf{Z}\mathbf{W}_N\hat{\mathbf{S}}\mathbf{W}_N\mathbf{Z}'\mathbf{X}][\mathbf{X}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\mathbf{X}]^{-1} \tag{6.97}$$

其中：

$$\hat{\mathbf{S}}=\frac{1}{N}\sum_{i=1}^{N}\mathbf{Z}_i'\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'\mathbf{Z}_i \tag{6.98}$$

需要特别注意对加权矩阵的几种不同选择。

第一，最优系统广义矩方法估计量(**optimal systems GMM estimator**)是满足 $\mathbf{W}_N=\hat{\mathbf{S}}^{-1}$ 的式(6.96)，其中，$\hat{\mathbf{S}}$ 已由式(6.98)定义。于是，其方差矩阵简化成：

$$\hat{\mathrm{V}}[\hat{\boldsymbol{\beta}}_{OSGMM}]=N\,[\mathbf{X}'\mathbf{Z}\hat{\mathbf{S}}^{-1}\mathbf{Z}'\mathbf{X}]^{-1}$$

该估计量是建立在矩条件(6.94)基础上的最有效的广义矩方法估计量。提高有效性源于两个因素：(1) 一个因素是系统估计，它允许不同方程的误差项是相关的，因此，$\mathrm{V}[\mathbf{u}_i|\mathbf{Z}_i]$并没有被限制为分块对角的；(2) 另一个因素是考虑相当一般的异方差性与相关性，故 $\boldsymbol{\Omega}_i$ 能随不同 i 而变化。

第二，当 $\mathbf{W}_N=(N^{-1}\mathbf{Z}'\mathbf{Z})^{-1}$ 时，得到系统 2SLS 估计量(**systems 2SLS estimator**)。考察由式(6.91)定义的看似不相关回归模型，一些回归元 $\mathbf{x}_{ig}$ 是内生的。假如我们定义工具矩阵是：

$$\mathbf{Z}_i=\begin{bmatrix}\mathbf{z}_{i1}' & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \ddots & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{z}_{iG}'\end{bmatrix} \tag{6.99}$$

系统 2SLS 就简化为包含第 g 个方程的工具 $\mathbf{z}_g$ 的逐一方程 2SLS。在许多应用中，$\mathbf{z}_1=\mathbf{z}_2=\cdots=\mathbf{z}_g$，共同工具集合用于所有方程，但我们并不需要把分析限制在这种情况。如果我们定义 $\mathbf{Z}_i=[\mathbf{z}_{i1}\cdots\mathbf{z}_{iT}]$，对面板数据模型(6.92)来说，系统 2SLS 就会简化成混合 2SLS。

第三，假定 $\mathrm{V}[\mathbf{u}_i|\mathbf{Z}_i]$并不随不同 i 而变化，因而 $\mathrm{V}[\mathbf{u}_i|\mathbf{Z}_i]=\boldsymbol{\Omega}$。这是一个与单方程异方差性假设类似的形式。于是，如同式(6.88)，$\boldsymbol{\Omega}$ 的一致估计值是 $\hat{\boldsymbol{\Omega}}=N^{-1}\sum_i\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'$，其中，$\hat{\mathbf{u}}_i$ 表示建立在一致工具变量估计量譬如系统 2SLS 基础上的残差。那么，最优广义矩方法估计量是满足 $\mathbf{W}_N=\mathbf{I}_N\otimes\hat{\boldsymbol{\Omega}}$ 的式(6.96)。该估计量应与下一节末尾所述的三阶段最小二乘法加以比较。

6.9.6 线性联立方程组

2.4 节引进的线性联立方程模型是一种非常重要的模型，研究生水平的经济计量学导论经常对此类模型详细阐述。本节给出一个非常简洁而又完整的概述。有关识别的讨论和第 2 章的内容相重叠。由于存在内生变量，所以 OLS 与 SUR 估计量均是非一致的。一些标准方法已在广义矩方法出现之前就得到很好的发展，但是，一致估计方法仍可放在广义矩方法估计的背景下。

线性联立方程模型设定如下，对于 N 个个体的第 i 个而言，G 个方程的第 g 个由下式给出：

$$y_{ig}=\mathbf{z}_{ig}'\gamma_g+\mathbf{Y}_{ig}'\beta_g+u_{ig}, \quad g=1,\cdots,G \tag{6.100}$$

其中，下标次序采用 6.9 节而不是 2.4 节的次序，$\mathbf{z}_g$ 表示外生回归元向量，假定外生回归元与误差项 u_g 是不相关的，而 $\mathbf{Y}_g$ 表示包括其他 $G-1$ 个方程的因变量 $y_{i1},\cdots, y_{g-1}, y_{g+1},\cdots, y_G$ 的子集向量。由于 $\mathbf{Y}_g$ 与模型误差相关，故 $\mathbf{Y}_g$ 是内生的。第 i 个个体的模型等价地写成：

$$\mathbf{y}_i'\mathbf{B}+\mathbf{z}_i'\mathbf{\Gamma}=\mathbf{u}_i \tag{6.101}$$

其中，$\mathbf{y}_i=[y_{i1}\cdots y_{iG}]'$ 表示 $G\times 1$ 维内生变量向量，$\mathbf{z}_i$ 表示 $r\times 1$ 维外生变量向量，$\mathbf{z}_i$ 是 $\mathbf{z}_{i1},\cdots,z_{iG}$ 的并集，$\mathbf{u}_i=[u_{i1}\cdots u_{iG}]'$ 表示 $G\times 1$ 维误差向量，$\mathbf{B}$ 表示 $G\times G$ 阶对角元素为 1 的参数矩阵，$\mathbf{\Gamma}$ 表示 $r\times G$ 阶参数矩阵，$\mathbf{B}$ 与 $\mathbf{\Gamma}$ 的一些元素被限制成 1。假定 $\mathbf{u}_i$ 对不同 i 是 iid 的，其均值为 $\mathbf{0}$，且方差矩阵为 $\mathbf{\Sigma}$。

把模型(6.101)称为结构式(**structural form**)，对应于各种不同结构，对 $\mathbf{B}$ 与 $\mathbf{\Gamma}$ 具有不同的限制。把内生变量求解为外生变量的函数，就得到简化式[1](**reduced form**)：

$$\begin{aligned}\mathbf{y}_i &= -\mathbf{z}_i'\mathbf{\Gamma}\mathbf{B}^{-1}+\mathbf{u}_i\mathbf{B}^{-1} \\ &=\mathbf{z}_i'\mathbf{\Pi}+\mathbf{v}_i\end{aligned} \tag{6.102}$$

其中，$\mathbf{\Pi}=-\mathbf{\Gamma}\mathbf{B}^{-1}$ 表示 $r\times G$ 阶简化式的参数矩阵，而 $\mathbf{v}_i=\mathbf{u}_i\mathbf{B}^{-1}$ 表示简化式误差向量，其方差为 $\mathbf{\Omega}=(\mathbf{B}^{-1})'\mathbf{\Sigma}\mathbf{B}^{-1}$。

简化式可通过 OLS 一致估计出来，得出 $\mathbf{\Pi}=-\mathbf{\Gamma}\mathbf{B}^{-1}$ 与 $\mathbf{\Omega}=(\mathbf{B}^{-1})'\mathbf{\Sigma}\mathbf{B}^{-1}$ 的估计值。如 2.5 节所示，识别问题意指，上述估计能否得出结构式参数 $\mathbf{B}$、$\mathbf{\Gamma}$ 以及 $\mathbf{B}^{-1}$ 的唯一估计值。由于对 $\mathbf{B}$、$\mathbf{\Gamma}$ 没有限制，而且 $\mathbf{\Sigma}$ 包含比 $\mathbf{\Pi}$ 与 $\mathbf{\Omega}$ 更多的参数，因此需要一些参数元素。第 g 个方程参数识别(**identification of parameters**)的必要条件是阶条件(**order condition**)，即第 g 个方程没有包含的外生变量数量必须至少等于包含的内生变量数量。这与 6.4.1 节给出的阶条件是一致的。例如，假如式(6.100)中的 $\mathbf{Y}_{ig}$ 具有一个元素，所以该方程存在一个内生变量，则 $\mathbf{x}_i$ 至少有一个元素必须没有被包含进来。这样做就保证存在与回归元一样多的工具。识别的充分条件是较强的秩条件。一些书籍，比如格林(Greene, 2003)等，都给出秩条件，

[1] 又称为简化型。——译者注

这里为了简洁起见不再阐述。其他一些约束,比如协方差约束,也会导致识别。

倘若已知识别,则通过式(6.44)定义的二阶段最小二乘法(**two-stage least squares**),对每个方程单独估计,就能一致估计出结构模型参数。同样的工具集合 $\mathbf{z}_i$ 可用于每一个方程。在第 g 个方程中,子元素 $\mathbf{z}_{ig}$ 用作其自身的工具,而 $\mathbf{z}_i$ 的其余元素则用作 $\mathbf{Y}_{ig}$ 的工具。

更有效的系统估计可以利用泽尔纳和泰尔(Zellner and Theil, 1962)的三阶段最小二乘法(3SLS)估计量来得到,该方法假定误差为同方差的,但对不同方程却是相关的。首先,通过 $\mathbf{y}$ 对 $\mathbf{z}$ 的 OLS 回归,估计出式(6.102)的简化系数 $\mathbf{\Pi}$。其次,通过式(6.100)的 OLS 回归,获得 2SLS 估计值,其中,$\mathbf{Y}_g$ 要用简化式预测值 $\hat{\mathbf{Y}}_g=\mathbf{z}'\hat{\mathbf{\Pi}}_g$ 代替。这正是 y_g 对 $\hat{\mathbf{Y}}_g$ 与 $\mathbf{z}_g$ 的 OLS 回归,或者等价地,y_g 对 $\hat{\mathbf{x}}_g$ 的 OLS 回归,其中,$\hat{\mathbf{x}}_g$ 表示来自关于 $\mathbf{z}$ 的 OLS 回归的对 $\mathbf{Y}_g$ 与 $\mathbf{z}$ 的预测。最后,通过 y_g 对 $\hat{\mathbf{x}}_g$ 的系统 OLS 回归,得出 3SLS 估计值,$g=1,\cdots,G$。于是,由式(6.89)可得出:

$$\hat{\boldsymbol{\theta}}_{3SLS}=[\hat{\mathbf{X}}'(\hat{\mathbf{\Sigma}}^{-1}\otimes\mathbf{I}_N)\hat{\mathbf{X}}]^{-1}\hat{\mathbf{X}}'(\hat{\mathbf{\Sigma}}^{-1}\otimes\mathbf{I}_N)\mathbf{y}$$

其中,$\hat{\mathbf{X}}$ 是通过首先建立具有对角分块 $\hat{\mathbf{x}}_{i1},\cdots,\hat{\mathbf{x}}_{iG}$ 的分块对角矩阵 $\hat{\mathbf{X}}_i$,然后叠放 $\hat{\mathbf{X}}_1,\cdots,\hat{\mathbf{X}}_N$ 而得到的,而具有 $\hat{\mathbf{u}}_i$ 残差向量的 $\hat{\mathbf{\Sigma}}=N^{-1}\sum_i\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'$ 则由 2SLS 估计值计算出。

在系统广义矩方法估计量的每一个方程均使用相同工具的情况下,这个估计量与满足 $\mathbf{W}_N=\mathbf{I}_N\otimes\hat{\mathbf{\Sigma}}$ 的系统广义矩方法估计量是一致的。另外,如果 $\mathrm{E}[\mathbf{u}_i|\mathbf{z}_i]=\mathbf{0}$,尽管 3SLS 与系统广义矩方法会产生一致估计值,但它们还是不同的。

6.9.7 线性方程组 ML 估计

本质上讲,线性模型的系统估计量是将推断建立在稳健标准误差基础上的 LS 或工具变量估计量。现在,还假定有正态分布 iid 的误差项,故 $\mathbf{u}_i\sim\mathcal{N}[\mathbf{0},\mathbf{\Omega}]$。

对具有外生回归元的方程组来说,所得到的 MLE 渐近地等价于 GLS 估计量。不过,这些估计量使用 $\mathbf{\Omega}$ 的不同估计量,从而 β 不同,因此,MLE 与 GLS 估计量之间存在小样本差异。例如,参见第 21 章的随机效应面板数据模型。

对线性联立方程组(6.101)来说,有限信息极大似然(**limited information maximum likelihood**)估计量,即单方程极大似然估计量,渐近地等价于 2SLS。完全信息极大似然(**full information maximum likelihood**)估计量,即系统 MLE,渐近地等价于 3SLS。例如,参见施密特(Schmidt, 1976)和格林(Greene, 2003)。

6.10 非线性方程组

现在,考察关于参数为非线性的方程组。例如,从所设定的直接效用或间接效用中得到的需求方程组,可能关于参数是非线性的。更一般地讲,如果非线性模型适用于孤立研究的因变量,譬如 logit 或泊松模型,那么这两个或更多变量的任何联立模型将一定是非线性的。

在关注偏参数建模之前,以对完全参数联立建模开始讨论。如同线性情况一

样，在考虑内生回归元复杂情况之前，阐述具有外生回归元的模型。

6.10.1 非线性方程组极大似然估计

5.6 节已经阐述单个因变量的极大似然估计。这些结果能立即应用到几个因变量的联立模型，只需要做出很小改变而已，即单个因变量的条件密度 $f(y_i|\mathbf{x}_i,\boldsymbol{\theta})$ 变为 $f(\mathbf{y}_i|\mathbf{X}_i,\boldsymbol{\theta})$，其中，$\mathbf{y}_i$ 表示因变量向量，$\mathbf{X}_i$ 表示所有回归元，而 $\boldsymbol{\theta}$ 表示所有参数。

例如，若 $y_1\sim\mathcal{N}[\exp(\mathbf{x}_1'\boldsymbol{\beta}_1),\ \sigma_1^2]$ 且 $y_2\sim\mathcal{N}[\exp(\mathbf{x}_2'\boldsymbol{\beta}_2),\ \sigma_2^2]$，则可假定适当的联立模型 $(y_1,\ y_2)$ 服从二元正态分布，其均值分别为 $\exp(\mathbf{x}_1'\boldsymbol{\beta}_1)$ 与 $\exp(\mathbf{x}_2'\boldsymbol{\beta}_2)$，方差分别为 σ_1^2 与 σ_2^2，且相关系数为 ρ。

对于数据不服从正态分布的情况，在设定和选取充分灵活联合分布的方面存在一些挑战。例如，对单变量计数而言，标准的开始模型是负二项式（参见第 20 章）。不过，在把这种方法推广到二变量或者多变量的计数模型时，存在几种可供选择的二变量负二项式模型。例如，就单变量条件分布或者单变量的边际分布是否是负二项式而论，它们是一样的。与之相比，多变量正态分布具有条件分布和边际分布，它们都服从正态分布。所有这些多变量负二项式对相关范围设置了某些约束，诸如正相关约束，但是对多变量正态而言，则不存在这种约束。

幸运的是，现代计算发展允许设定较为丰富的模型。例如，假定一种合情合理的相关二变量计数的灵活模型，是以不可观测值 ε_1 与 ε_2 为条件的，y_1 服从均值为 $\exp(\mathbf{x}_1'\boldsymbol{\beta}_1+\varepsilon_1)$ 的泊松分布，y_2 服从均值为 $\exp(\mathbf{x}_1'\boldsymbol{\beta}_1+\varepsilon_2)$ 的泊松分布。估计二变量分布可通过假定不可观测的 ε_1 与 ε_2 服从二变量正态分布，并且通过积分去掉 ε_1 与 ε_2 来获得。对这种二变量分布来说，不存在闭型解，但其参数却可利用 12.4 节将阐述的极大模拟似然法得到估计。

本书的第 4 部分将给出非线性联立模型的一系列例子。最简单的联立模型不具有灵活性，因而一致性依赖于约束非常强的分布假设。不过，一般地讲，设定更灵活的模型能利用计算密集方法加以估计，这样做不存在理论上的障碍。

特别地，19.3 节将详细阐述两种重要的方法，用于生成丰富的多变量参数模型。这些方法在持续期限数据模型的背景下给出，但具有更广泛的可应用性。首先，可引入相关的不可观测异质性（**unobserved heterogeneity**），如同在二变量计数例子中那样。其次，可以运用联接[1]（**copulas**），它提供一种生成已知设定单变量边缘分布的联合分布的方法。

对极大似然估计来说，一种比较简单却稍欠有效性的拟极大似然方法是，设定 y_1 与 y_2 的各自参数模型，并且通过 y_1 与 y_2 的独立性，可获得极大似然估计值，不过，允许 y_1 与 y_2 是相关的，就可实施统计推断。这些内容已由 5.7.5 节阐述。本节的余下部分考察此类偏参数方法。

倘若存在内生性，则存在较大的挑战性，因此，在一个方程中出现的因变量，作

---

〔1〕 这里把“copulas”译为“联接”，以便与一般广义模型中的另一个被称为标准连接（canonical link）的术语相区别。有人把“copulas”译成连接。——译者注

为另一个方程的回归元。除带有服从正态分布可加误差的非线性回归模型之外，存在极少数非线性联立方程。

6.10.2 *非线性方程组*

就线性回归而言，从单方程到多个方程的变动是显而易见的，因为起点是线性模型 $y=\mathbf{x}'\boldsymbol{\beta}+u$，并利用最小二乘法估计，而有效系统估计则利用系统 GLS 来获得。对于非线性模型，无论是研究起点还是估计方法，都存在相当大的差异，且研究手段各异。

将含有 G 个因变量的多变量非线性模型（**multivariate nonlinear model**）定义为：

$$\mathbf{r}(\mathbf{y}_i,\mathbf{X}_i,\boldsymbol{\beta})=\mathbf{u}_i \tag{6.103}$$

其中，$\mathbf{y}_i$ 与 $\mathbf{u}_i$ 表示 $G\times1$ 维向量，$\mathbf{r}(\mathbf{y}_i,\mathbf{X}_i,\boldsymbol{\beta})$ 表示 $G\times1$ 维向量函数，$\mathbf{X}_i$ 表示 $G\times L$ 阶矩阵，而 $\boldsymbol{\beta}$ 表示 $K\times1$ 维列向量。本节中，我们自始至终地做出横截面假设：误差向量 $\mathbf{u}_i$ 对不同 i 是独立的，但是，给定 i 时，$\mathbf{u}_i$ 元素可能与随 i 变化的方差及协方差相关。

式(6.103)的一个例子是非线性看似不相关回归模型（**nonlinear seemingly unrelated regression model**）。于是，对 N 个个体的第 i 个而言，G 个方程的第 g 个方程由下式给出：

$$r_g(y_{ig},\mathbf{x}_{ig},\boldsymbol{\beta}_g)=u_{ig},\quad g=1,\cdots,G \tag{6.104}$$

例如，$u_{ig}=y_{ig}-\exp(\mathbf{x}_{ig}'\boldsymbol{\beta}_g)$。于是，式(6.103)的 $\mathbf{u}_i$ 与 $\mathbf{r}(\cdot)$ 表示 $G\times1$ 维向量，其第 g 个元素为 u_{ig} 与 $r_g(\cdot)$，$\mathbf{X}_i$ 表示与式(6.91)所定义的矩阵相同的分块对角矩阵，而 $\boldsymbol{\beta}$ 是通过把 $\boldsymbol{\beta}_1$ 至 $\boldsymbol{\beta}_G$ 叠放得到的。

第二个例子是非线性面板数据模型（**nonlinear panel data model**）。于是，对处于时期 t 的个体 i 来说，有：

$$r(y_{it},\mathbf{x}_{it},\boldsymbol{\beta})=u_{it},\quad t=1,\cdots,T \tag{6.105}$$

从而，式(6.103)中的 $\mathbf{u}_i$ 与 $\mathbf{r}(\cdot)$ 表示 $T\times1$ 维向量，因而 $G=T$，其第 t 个元素为 u_{it} 与 $r(y_{it},\mathbf{x}_{it},\boldsymbol{\beta})$。这种面板模型不同于在每一个时期都拥有同样函数 $r(\cdot)$ 与参数 $\boldsymbol{\beta}$ 的看似不相关回归模型。

6.10.3 *非线性系统估计*

当模型(6.103)的回归元 $\mathbf{X}_i$ 都是外生的时，有：

$$\mathrm{E}[\mathbf{u}_i|\mathbf{X}_i]=\mathbf{0} \tag{6.106}$$

其中，$\mathbf{u}_i$ 表示模型(6.103)定义的误差项。我们假定误差项对于不同 i 是独立的，且方差矩阵是：

$$\boldsymbol{\Omega}_i=\mathrm{E}[\mathbf{u}_i\mathbf{u}_i'|\mathbf{X}_i] \tag{6.107}$$

可加误差

当非线性模型关于误差项是可加的时候，系统估计就是对线性模型的系统OLS与FGLS估计的直接改进，因此，对式(6.103)进行专门研究，得出：

$$\mathbf{u}_i=\mathbf{y}_i-\mathbf{g}(\mathbf{X}_i,\boldsymbol{\beta}) \tag{6.108}$$

于是，系统NLS估计量(**systems NLS estimator**)对残差平方和$\sum_i\mathbf{u}_i'\mathbf{u}_i$求极小值，而系统FGNLS估计量(**systems FGNLS estimator**)则对：

$$Q_N(\boldsymbol{\beta})=\sum_i\mathbf{u}_i'\hat{\boldsymbol{\Omega}}_i^{-1}\mathbf{u}_i \tag{6.109}$$

求极小值，其中，将$\boldsymbol{\Omega}_i$设定成模型$\boldsymbol{\Omega}_i(\gamma)$，并且$\hat{\boldsymbol{\Omega}}_i=\boldsymbol{\Omega}_i(\hat{\gamma})$。为了防止对$\boldsymbol{\Omega}_i$可能错误的设定，人们能使用本质上仅要求$\mathbf{u}_i$是独立的且满足式(6.106)的稳健标准误差。于是，系统FGNLS估计量的估计方差与式(6.87)中线性系统FGLS的估计方差是一样的，只是用$\partial\mathbf{g}(\mathbf{y}_i,\boldsymbol{\beta})/\partial\boldsymbol{\beta}'|_{\hat{\boldsymbol{\beta}}}$代替$\mathbf{X}_i$，现在有$\hat{\mathbf{u}}_i=\mathbf{y}_i-\mathbf{g}(\mathbf{X}_i,\hat{\boldsymbol{\beta}})$。比较简单系统NLS估计量的估计方差可通过另外用$\mathbf{I}_G$代替$\hat{\boldsymbol{\Omega}}_i$来得到。

主要挑战是对$\boldsymbol{\Omega}_i$设定一个有用模型。举一个例子，假定要对两个计数变量进行联合建模。第20章将证明，一种比泊松分布模型稍欠一般性的标准计数模型是，将条件均值设定为$\exp(\mathbf{x}'\boldsymbol{\beta})$，同时把条件方差设定为$\exp(\mathbf{x}'\boldsymbol{\beta})$。然后，将联合模型设定为$\mathbf{u}=[u_1\quad u_2]'$，其中，$u_1=y_1-\exp(\mathbf{x}_1'\boldsymbol{\beta}_1)$，$u_2=y_2-\exp(\mathbf{x}_2'\boldsymbol{\beta}_2)$。于是，方差矩阵$\boldsymbol{\Omega}_i$具有对角元素$\alpha_1\exp(\mathbf{x}_{i1}'\boldsymbol{\beta}_1)$与$\alpha_2\exp(\mathbf{x}_{i2}'\boldsymbol{\beta}_2)$，而且对协方差的一种可能参数化是$\alpha_3[\exp(\mathbf{x}_{i1}'\boldsymbol{\beta}_1)\exp(\mathbf{x}_{i2}'\boldsymbol{\beta}_2)]^{1/2}$。因此，要估计$\hat{\boldsymbol{\Omega}}_i$，就需要$\boldsymbol{\beta}_1$、$\boldsymbol{\beta}_2$、$\alpha_1$、$\alpha_2$以及$\alpha_3$的值，而这些值可从第一步单方程估计中获得。

非可加误差

正如6.2.2节在单方程情况下所证明的，对非可加误差而言，最小二乘法回归已不再适用。伍德里奇(Wooldridge, 2002)阐述了矩估计的一致方法。

条件矩约束(6.106)产生许多能用于估计的可行无条件的矩条件。一个明显的起点是，把估计建立在矩条件$\mathrm{E}[\mathbf{X}_i'\mathbf{u}_i]=\mathbf{0}$的基础上。不过，可使用另外矩条件。一般地讲，考察建立在K个矩条件：

$$\mathrm{E}[\mathbf{R}(\mathbf{X}_i,\boldsymbol{\beta})\mathbf{u}_i]=\mathbf{0} \tag{6.110}$$

基础上的估计，其中，$\mathbf{R}(\mathbf{X}_i,\boldsymbol{\beta})$表示$\mathbf{X}_i$与$\boldsymbol{\beta}$的$K\times G$阶矩阵函数。对$\mathbf{R}(\mathbf{X}_i,\boldsymbol{\beta})$进行设定，而且可能依赖于$\boldsymbol{\beta}$，这一点将在下面加以讨论。

由构造可知，存在与参数同样多的矩条件。系统矩方法估计量(**system method of moment estimator**)$\hat{\boldsymbol{\beta}}_{\mathrm{SMM}}$求解相应的样本矩条件：

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{R}(\mathbf{X}_i,\boldsymbol{\beta})'\mathbf{r}(\mathbf{y}_i,\mathbf{X}_i,\hat{\boldsymbol{\beta}}_{\mathrm{SMM}})=\mathbf{0} \tag{6.111}$$

其中，$\mathbf{R}(\mathbf{X}_i,\boldsymbol{\beta})$实际上是在第一步估计$\tilde{\boldsymbol{\beta}}$时计算出来的。该估计量服从渐近正态分布，其方差矩阵为：

$$\hat{\mathrm{V}}[\hat{\boldsymbol{\beta}}_{\mathrm{SMM}}]=\Big[\sum_{i=1}^{N}\hat{\mathbf{D}}_i'\hat{\mathbf{R}}_i\Big]^{-1}\sum_{i=1}^{N}\hat{\mathbf{R}}_i'\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'\hat{\mathbf{R}}_i\Big[\sum_{i=1}^{N}\hat{\mathbf{R}}_i'\hat{\mathbf{D}}_i\Big]^{-1} \tag{6.112}$$

其中，$\hat{\mathbf{D}}_i=\partial\mathbf{r}_i/\partial\boldsymbol{\beta}'|_{\hat{\boldsymbol{\beta}}}$，$\hat{\mathbf{R}}_i=\mathbf{R}(\mathbf{X}_i,\hat{\boldsymbol{\beta}})$，而$\hat{\mathbf{u}}_i=\mathbf{r}(\mathbf{y}_i,\mathbf{X}_i,\hat{\boldsymbol{\beta}}_{\mathrm{SMM}})$。

主要问题是对式(6.110)中的 $\mathbf{R}(\mathbf{X},\boldsymbol{\beta})$ 进行设定。由 6.3.7 节知，建立在式(6.106)基础上的最有效估计量设定：

$$\mathbf{R}^*(\mathbf{X}_i,\boldsymbol{\beta})=\mathrm{E}\left[\frac{\partial\mathbf{r}(\mathbf{y}_i,\mathbf{X}_i,\boldsymbol{\beta})'}{\partial\boldsymbol{\beta}}\Big|\mathbf{X}_i\right]\boldsymbol{\Omega}_i^{-1} \tag{6.113}$$

通常，如果得出最优估计比较困难，则右边第一个表达式需要强分布假设。

不过，若非线性模型具有式(6.108)定义的可加误差，则可进行简化。于是，$\mathbf{R}^*(\mathbf{X}_i,\boldsymbol{\beta})=\partial\mathbf{g}(\mathbf{X}_i,\boldsymbol{\beta})'/\partial\boldsymbol{\beta}\times\boldsymbol{\Omega}_i^{-1}$，并且估计方程(6.110)变成：

$$N^{-1}\sum_{i=1}^{N}\frac{\partial\mathbf{g}(\mathbf{X}_i,\boldsymbol{\beta})'}{\partial\boldsymbol{\beta}}\boldsymbol{\Omega}_i^{-1}(\mathbf{y}_i-\mathbf{X}_i'\hat{\boldsymbol{\beta}}_{\mathrm{SMM}})=\mathbf{0}$$

此估计量渐近地等价于对式(6.109)求极小值的系统 FGNLS 估计量。

6.10.4 *非线性系统工具估计*

当模型(6.103)中的回归元 $\mathbf{X}_i$ 是内生的时，$\mathrm{E}[\mathbf{u}_i|\mathbf{X}_i]\neq\mathbf{0}$，假定存在 $G\times r$ 阶工具矩阵 $\mathbf{Z}_i$，使得：

$$\mathrm{E}[\mathbf{u}_i|\mathbf{Z}_i]=\mathbf{0} \tag{6.114}$$

其中，$\mathbf{u}_i$ 表示模型(6.103)定义的误差项。我们假定误差项对于不同 i 是独立的，且方差矩阵是 $\boldsymbol{\Omega}_i=\mathrm{E}[\mathbf{u}_i\mathbf{u}_i'|\mathbf{Z}_i]$。对非线性 SUR 模型来说，$\mathbf{Z}_i$ 如同式(6.99)所定义的。

这个方法类似于前一节关于系统矩方法估计量的那种方法，具有额外的复杂性，即可能存在着剩余工具导致需要广义矩方法估计而不是矩方法估计的情况。条件矩约束(6.106)会产生许多用于估计的无条件矩条件。此外，我们遵从把估计建立在矩条件 $\mathrm{E}[\mathbf{Z}_i'\mathbf{u}_i]=\mathbf{0}$ 基础上的许多其他线索。那么，系统广义矩方法估计量(**systems GMM estimator**)对

$$Q_N(\beta)=\Big[\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{r}(\mathbf{y}_i,\mathbf{X}_i,\boldsymbol{\beta})\Big]'\mathbf{W}_N\Big[\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{r}(\mathbf{y}_i,\mathbf{X}_i,\boldsymbol{\beta})\Big] \tag{6.115}$$

求极小值。此估计量服从渐近正态分布，其估计方差为：

$$\hat{\mathrm{V}}[\hat{\beta}_{\mathrm{SGMM}}]=N[\hat{\mathbf{D}}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\hat{\mathbf{D}}]^{-1}[\hat{\mathbf{D}}'\mathbf{Z}\mathbf{W}_N\hat{\mathbf{S}}\mathbf{W}_N\mathbf{Z}'\hat{\mathbf{D}}][\hat{\mathbf{D}}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\hat{\mathbf{D}}]^{-1} \tag{6.116}$$

其中，$\hat{\mathbf{D}}'\mathbf{Z}=\sum_i\partial\mathbf{r}_i'/\partial\boldsymbol{\beta}|_{\hat{\beta}}\mathbf{Z}_i$，且 $\hat{\mathbf{S}}=N^{-1}\sum_i\mathbf{Z}_i\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'\mathbf{Z}_i'$，同时假定 $\mathbf{u}_i$ 对不同 i 是独立的，其方差矩阵 $\mathrm{V}[\mathbf{u}_i|\mathbf{X}_i]=\boldsymbol{\Omega}_i$。

在从非线性看似不相关回归模型中得到 $\mathbf{r}(\mathbf{y}_i,\mathbf{X}_i,\boldsymbol{\beta})$ 的情况下，选择 $\mathbf{W}_N=[N^{-1}\sum_i\mathbf{Z}_i\mathbf{Z}_i']^{-1}$ 对应于 NL2SLS。将选择 $\mathbf{W}_N=[N^{-1}\sum_i\mathbf{Z}_i\hat{\boldsymbol{\Omega}}\mathbf{Z}_i']^{-1}$ 称为非线性 3SLS(**nonlinear 3SLS**，缩记为 NL3SLS)，并且是在 $\boldsymbol{\Omega}_i=\boldsymbol{\Omega}$ 的特殊情况下、建立在矩条件 $\mathrm{E}[\mathbf{Z}_i'\mathbf{u}_i]=\mathbf{0}$ 基础上的最有效估计量，其中，$\hat{\boldsymbol{\Omega}}=N^{-1}\sum_i\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'$。在更一般的假设下，即 $\boldsymbol{\Omega}_i$ 随不同 i 而变化，选择 $\mathbf{W}_N=\hat{\mathbf{S}}^{-1}$ 会得出最有效估计量。不过，与以往一样，通过矩条件而不是 $\mathrm{E}[\mathbf{Z}_i'\mathbf{u}_i]=\mathbf{0}$，会得出更有效的估计量。

6.10.5 *非线性联立方程组*

非线性联立方程模型(**nonlinear simultaneous equations model**)设定如下，对于

N 个个体的第 i 个，设定 G 个方程的第 g 个由下式给出：

$$u_{ig}=r_g(\mathbf{y}_i,\mathbf{x}_{ig},\boldsymbol{\beta}_g),\quad g=1,\cdots,G \tag{6.117}$$

这是具有包括来自其他方程因变量的回归元的非线性看似不相关回归模型。与线性联立方程模型不同，从应用上看，确保非线性联立方程模型是可识别的结果很少。

已知识别，利用前一节阐述的广义矩方法估计量获得一致估计值。或者，我们假定 $\mathbf{u}_i\sim\mathcal{N}[\mathbf{0},\boldsymbol{\Omega}]$，从而得到非线性完全信息极大似然估计量（**nonlinear full-information maximum likelihood estimator**）。对背离线性联立方程模型的情况来说，通常非线性完全信息 MLE 具有不同于 NL3SLS 的渐近分布，同时非线性完全信息极大似然估计要求，误差项在本质上服从正态分布。有关详细内容，参见雨宫(Amemiya, 1985)。

处置非线性模型的内生性极为复杂。16.8 节将考察 Tobit 模型中的联立性(simultaneity)，当模型关于潜变量是线性的时候，分析起来比较简单。20.6.2 节考察一种高度非线性的例子，即计数数据模型中的内生回归元。

6.11 应用研究

理想上，利用经济计量软件包能实施广义矩方法，这样就不会遇到更多的困难，也不会需要更多的知识，例如，具有异方差误差的非线性最小二乘法估计。然而，不是所有的重要经济计量软件包都能提供广泛的广义矩方法模块。依据特定应用，需要将广义矩方法估计转换成一种更合适的软件包，或者使用具有广义矩方法代数运算的矩阵程序。

广义矩方法的一种普遍应用是工具变量估计。大多数经济计量学软件包涉及线性工具变量估计量，但不是所有的软件包都涵盖非线性工具变量估计量。默认标准误差可能假定同方差误差，而不是异方差稳健的。正如第 4 章强调的，很难获得与误差项不相关但与回归元非常相关的工具，或者，在非线性情况条件下，很难获得有关参数的对误差的适当推导。

经济计量学软件包通常包含线性方程组，却不包含非线性方程组。并且，默认标准误差对异方差性来说，可能不是稳健的。

6.12 文献注释

对广义矩方法进行研究的教科书，包括戴维森和麦金龙(Davidson and MacKinnon, 1993, 2004)、哈密尔顿(Hamilton, 2004)以及格林(Green, 2003)。最近，由林文夫(Hayashi, 2005)与伍德里奇(Wooldridge, 2002)所撰写的书特别强调广义矩方法估计。贝拉和比林阿斯(Bera and Bilias, 2002)给出本书第 5 章和第 6 章阐述的一些估计量的综述及历史。

6.3 广义矩方法的原创性文献是汉森(Hansen, 1982)的论文。阿雷拉诺

(Arellano, 2003)的《面板数据的经济计量学》附录,给出关于广义矩方法最优矩的一个很好的解释。《商业和经济统计学杂志》(*Journal of Business and Economic Statistics*)2002 年 10 月份专刊致力于 GMM 估计。

6.4 萨根(Sargan, 1958)对线性工具变量估计的经典研究是广义矩方法的重要前身。

6.5 由雨宫(Amemiya, 1974)引入的非线性 2SLS 估计量,很容易被推广到广义矩方法估计量。

6.6 时序两阶段估计的标准参考文献是纽韦(Newey, 1984)、托佩尔和墨菲(Murphy and Topel, 1985)以及帕甘(Pagan, 1986)的论文。

6.7 最小距离估计的标准参考文献是张伯伦(Chamberlain, 1982)的论文。

6.8 对经验似然估计做出的良好概述,由米特尔哈默、贾奇和米勒(Mittelhammer, Judge and Miller, 2000)提供,重要参考文献是欧文(Owen, 1988, 2001)以及勤和劳利斯(Qin and Lawless, 1994)的论文。英伯斯(Imbens, 2002)给出这种相对新颖方法的评论及应用。

6.9 例如,格林(Greene, 2003)的教科书提供了比此处内容更详细的关于系统估计的概述,特别是关于线性看似不相关回归与线性联立方程模型。

6.10 雨宫(Amemiya, 1985)详细地阐述了非线性联立方程。

习 题

6-1 考察习题 5.2 的伽玛回归模型,有 $\mathrm{E}[y|\mathbf{x}]=\exp(\mathbf{x}'\boldsymbol{\beta})$ 且 $\mathrm{V}[y|\mathbf{x}]=(\exp(\mathbf{x}'\boldsymbol{\beta}))^2/2$。

(a) 证明这些条件蕴含 $\mathrm{E}[\mathbf{x}\{(y-\exp(\mathbf{x}'\boldsymbol{\beta}))^2-(\exp(\mathbf{x}'\boldsymbol{\beta}))^2/2\}]=\mathbf{0}$。

(b) 使用(a)部分的矩条件,建立矩方法估计量 $\hat{\boldsymbol{\beta}}_{\mathrm{MM}}$。

(c) 利用结果(6.13),给出 $\hat{\boldsymbol{\beta}}_{\mathrm{MM}}$ 的渐近分布。

(d) 除(a)部分之外,假定还可以利用矩条件 $\mathrm{E}[\mathbf{x}(y-\exp(\mathbf{x}'\boldsymbol{\beta}))]=0$。给出 $\boldsymbol{\beta}$ 的广义矩方法估计量的目标函数。

6-2 考察对于不同 i,数据独立的线性回归模型,$y_i=\mathbf{x}_i'\boldsymbol{\beta}+u_i$。假定 $\mathrm{E}[u_i|\mathbf{x}_i]\neq 0$,但存在着满足 $\mathrm{E}[u_i|\mathbf{z}_i]=0$ 且 $\mathrm{V}[u_i|\mathbf{z}_i]=\sigma_i^2$ 的可利用的工具 $\mathbf{z}_i$,其中,$\dim(\mathbf{z})>\dim(\mathbf{x})$。求极小化

$$Q_N(\boldsymbol{\beta})=\left[N^{-1}\sum_i \mathbf{z}_i(y_i-\mathbf{x}_i'\boldsymbol{\beta})\right]'\mathbf{W}_N\left[N^{-1}\sum_i \mathbf{z}_i(y_i-\mathbf{x}_i'\boldsymbol{\beta})\right]$$

的广义矩方法估计量。

(a) 利用一般的广义矩方法结果(6.11),推导 $\sqrt{N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)$ 的极限分布。

(b) 阐述如何获得 $\hat{\boldsymbol{\beta}}$ 的渐近方差的一致估计值。

(c) 如果误差是同方差的,你会使用 $\mathbf{W}_N$ 的哪种选择?请解释你的解答。

(d) 如果误差是异方差的,你会使用 $\mathbf{W}_N$ 的哪种选取?请解释你的解答。

6-3 考察 6.3.6 节末尾处拉普拉斯唯一截距的例子,因而有 $y=\mu+u$。于是,广义矩方法估计建立在 $\mathrm{E}[\mathbf{h}(\mu)]=\mathbf{0}$ 的基础上,其中,$\mathbf{h}(\mu)=[(y-\mu),(y-\mu)^3]'$。

(a) 利用 6.3.6 节给出的中心矩知识，证明 $\mathbf{G}_0=\mathrm{E}[\partial\mathbf{h}/\partial\mu]=[-1,-6]'$，并且 $\mathbf{S}_0=\mathrm{E}[\mathbf{hh}']$ 具有对角元素 2 与 720，而非对角元素为 24。

(b) 证明 $\mathbf{G}_0'\mathbf{S}_0^{-1}\mathbf{G}_0=252/432$。

(c) 证明 $\hat{\mu}_{\mathrm{OGMM}}$ 具有渐近方差 $1.7143/N$。

(d) 证明满足 $\mathbf{W}=\mathbf{I}_2$ 的 μ 的广义矩方法估计量具有渐近方差 $19.14/N$。

6-4 这个问题使用 probit 模型，但需要很少的模型知识。设 y 表示依据事件是否发生而取值为 0 或 1 的二值变量，设 $\mathbf{x}$ 表示回归元向量，并且假定独立观测值。

(a) 假定 $\mathrm{E}[y|\mathbf{x}]=\Phi(\mathbf{x}'\boldsymbol{\beta})$，其中，$\Phi(\cdot)$ 表示标准正态 cdf。证明 $\mathrm{E}[(y-\Phi(\mathbf{x}'\boldsymbol{\beta}))\mathbf{x}]=\mathbf{0}$。从而给出 $\boldsymbol{\beta}$ 的矩方法的估计方程。

(b) 这个估计量将会产生与 probit 极大似然估计相同的估计值吗？（对于这部分内容来说，你只需要阅读 14.3 节。）

(c) 给出(a)部分中的广义矩方法目标函数。也就是说，给出可产生相同一阶条件的目标函数，直到满秩的矩阵变换，就像在(a)部分获得的那样。

(d) 假定由于某些元素的内生性，$\mathrm{E}[y|\mathbf{x}]\neq\Phi(\mathbf{x}'\boldsymbol{\beta})$。假定存在一个向量 $\mathbf{z}$，$\dim[\mathbf{z}]>\dim[\mathbf{x}]$，使得 $\mathrm{E}[y-\Phi(\mathbf{x}'\boldsymbol{\beta})|\mathbf{z}]=0$。给出 $\boldsymbol{\beta}$ 的一致估计量的目标函数。此估计量并不需要是完全有效的。

(e) 对(d)部分所获得的估计量来说，给出其渐近分布。为了得到此结果，叙述你对数据生成过程做出的任何假设。

(f) 对(d)部分的最优广义矩方法估计量来说，给出加权矩阵，并且计算它的方法。

(g) 给出(d)部分的一个真实情况的例子。也就是说，给出具有内生回归元与有效工具的 probit 模型的一个有意义的例子。叙述用作允许进行一致估计的因变量、内生回归元以及工具。（令人意想不到的是，这部分很难。）

6-5 假定采用约束 $\mathrm{E}[\mathbf{w}_i]=\mathbf{g}(\theta)$，其中，$\dim[\mathbf{w}]>\dim[\boldsymbol{\theta}]$。

(a) 求广义矩方法估计量的目标函数。

(b) 求满足 $\boldsymbol{\pi}=\mathrm{E}[\mathbf{w}_i]$ 和 $\hat{\boldsymbol{\pi}}=\bar{w}$ 的最小距离估计量的目标函数（参见 6.7 节）。

(c) 证明在这个例子中，最小距离方法与广义矩方法是等价的。

6-6 最小距离估计量（参见 6.7 节）运用约束 $\boldsymbol{\pi}-\mathbf{g}(\boldsymbol{\theta})=\mathbf{0}$。更一般地讲，假定约束为 $\mathbf{h}(\boldsymbol{\theta},\boldsymbol{\pi})=\mathbf{0}$，并利用广义最小距离估计量来进行估计，即对 $Q_N(\boldsymbol{\theta})=\mathbf{h}(\boldsymbol{\theta},\hat{\boldsymbol{\pi}})'\mathbf{W}_N\mathbf{h}(\boldsymbol{\theta},\hat{\boldsymbol{\pi}})$ 求极小值。利用式(6.68)～(6.70)证明，当 $\mathbf{G}_0=\partial\mathbf{h}(\boldsymbol{\theta},\boldsymbol{\pi})/\partial\boldsymbol{\theta}|_{\theta_0,\pi_0}$ 且用 $\mathbf{H}_0'\mathrm{V}[\hat{\boldsymbol{\pi}}]\mathbf{H}_0$ 代替 $\mathrm{V}[\hat{\boldsymbol{\pi}}]$ 时，式(6.67)成立，其中 $\mathbf{H}_0=\partial\mathbf{h}(\boldsymbol{\theta},\boldsymbol{\pi})/\partial\pi|_{\theta_0,\pi_0}$。

6-7 利用 6.6.4 节给出的数据生成过程所生成的数据，$N=1\ 000$，求 NL2SLS 估计值，并且将这些值与两阶段估计值加以比较。



假设检验

7.1 引 论

本章考察参数可能为非线性的假设检验，其中利用了适合于非线性模型的估计量。

检验统计量的分布可利用与用于分析估计量的相同的统计理论来获得，这是因为像估计量一样的检验统计量仍然是一个统计量，也就是样本的函数。给定估计量与假设的适当线性化，其结果非常类似于线性回归模型对线性约束所进行的那些检验。不过，结果却依赖于渐近理论，同时在正态性下，线性模型的准确 t 分布与 F 分布的检验统计量要被用作渐近标准正态分布（z 检验）或卡方分布的检验统计量来代替。

在实施假设检验时，实际应用中存在两个重要的担心。首先，检验可能具有错误水平，因而在一种名义显著性譬如 5%上进行检验时，对零假设拒绝的真实概率很可能大于 5%。当基本渐近分布理论只是一种近似时，这种错误水平在适度水平的样本中几乎一定会产生。一种纠正方法是本章将引进的自助法，第 11 章将重点而广泛地对自助法加以研究。其次，一些检验可能具有低势[1]（low power），因此，当应该拒绝零假设时，却存在很小概率拒绝零假设。和大多数教科书对检验的研究相比，本章更加强调对水平与势的研究。

最广泛运用的检验方法[2]（procedure）即沃尔德检验由 7.2 节加以定义。另外，当利用极大似然估计时，7.3 节阐述似然比检验与得分检验，或者拉格朗日乘子检验。7.4 节对各种检验举例说明。7.5 节把这些检验扩展到估计量而不是极大似然上，包括稳健检验形式。7.6 节、7.7 节和 7.8 节则分别阐述检验势、蒙特卡罗模拟方法以及自助法。

第 8 章将独立地给出对模型设定以及选择的一些方法，而不涉及假设检验本身的研究。

---

〔1〕 又称为低功效。——译者注
〔2〕 又称为程序。——译者注

7.2 沃尔德检验

归功于沃尔德(Wald, 1943)的沃尔德检验,是微观经济计量学中一个极为出色的假设检验。它需要对无约束模型进行估计,也就是说,没有利用零假设约束的模型。沃尔德检验的应用相当广泛,因为人们通常利用现代软件对无约束模型进行估计,即使无约束模型比约束模型更为复杂,而且日益发展的现代软件提供了在相对弱分布的假设下允许沃尔德检验的稳健方差矩阵估计值。运用计算机软件报告出的对回归元统计显著性进行检验的通常统计量,就是沃尔德检验统计量的一个例子。

本节详细阐述非线性假设的沃尔德检验,既阐述理论又给出说明例子。并且,阐述与δ方法密切联系的、用于构建参数的非线性函数的置信区间或者置信区域的方法。本节末尾将详述沃尔德检验的弱点,即对于在代数形式上等价的零假设参数化来说,它缺乏不变性。

7.2.1 线性模型的线性假设

首先,回顾标准的线性模型结果,这是因为,沃尔德检验是对线性回归模型的线性约束进行通常检验的推广。

在线性回归模型$\mathbf{y}=\mathbf{X}'\boldsymbol{\beta}+\mathbf{u}$中,关于回归参数的线性约束的双侧检验的零假设与备选假设分别是:

$$\begin{aligned} H_0:&\ \mathbf{R}\boldsymbol{\beta}_0-\mathbf{r}=\mathbf{0}\\ H_a:&\ \mathbf{R}\boldsymbol{\beta}_0-\mathbf{r}\neq\mathbf{0}\end{aligned}\tag{7.1}$$

这里所用记号表示有h个约束,$\mathbf{R}$表示满秩h的$h\times K$阶常值矩阵,$\boldsymbol{\beta}$表示$K\times1$维参数向量,$\mathbf{r}$表示$h\times1$维常值向量,而且$h\leqslant K$。

例如,当$K=4$时,联合检验$\beta_1=1$和$\beta_2-\beta_3=2$能表述成式(7.1),满足:

$$\mathbf{R}=\begin{bmatrix}1&0&0&0\\0&1&-1&0\end{bmatrix},\quad \mathbf{r}=\begin{bmatrix}1\\2\end{bmatrix}$$

$\mathbf{R}\boldsymbol{\beta}_0-\mathbf{r}=\mathbf{0}$的沃尔德检验是对样本类似形式$\mathbf{R}\hat{\boldsymbol{\beta}}-\mathbf{r}$接近于0的检验,其中,$\hat{\boldsymbol{\beta}}$表示无约束OLS估计量。在$\mathbf{u}\sim\mathcal{N}[\mathbf{0},\sigma_0^2\mathbf{I}]$的强假设下,估计量$\hat{\boldsymbol{\beta}}\sim\mathcal{N}[\boldsymbol{\beta}_0,\sigma_0^2(\mathbf{X}'\mathbf{X})^{-1}]$,因此,在$H_0$下,有:

$$\mathbf{R}\hat{\boldsymbol{\beta}}-\mathbf{r}\sim\mathcal{N}[\mathbf{0},\sigma_0^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']$$

其中,$\mathbf{R}\boldsymbol{\beta}_0-\mathbf{r}=\mathbf{0}$被简化成$\mathbf{0}$的均值。当取二次形式时,得到检验统计量:

$$\mathrm{W}_1=(\mathbf{R}\hat{\boldsymbol{\beta}}-\mathbf{r})[\sigma_0^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}}-\mathbf{r})$$

在H_0下,该统计量确实服从$\chi^2(h)$分布。不过,在实际应用中,并不能计算出检验统计量W_1,因为σ_0^2是未知的。

在大样本中,σ_0^2用它的估计值s^2代替,并不会影响到W_1的极限分布,因为这

等价于用 σ_0^2/s^2 左乘 W_1，并且 $\text{plim}(\sigma_0^2/s^2)=1$（参见变换定理 A.12）。因此，下式

$$W_2=(\mathbf{R}\hat{\boldsymbol{\beta}}-r)'[s^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}}-r) \tag{7.2}$$

收敛到 H_0 下的 $\chi^2(h)$分布。

检验统计量 W_2 仅在渐近形式下服从卡方分布。在这种带有正态误差的线性例子中，能获得一种可供选择的准确的小样本结果。许多引论教科书所推导的标准结果是，在 H_0 下，若 $s^2=(N-K)^{-1}\sum_i \hat{u}_i$，则

$$W_3=W_2/h$$

服从 $F(h,N-K)$分布，其中，$\hat{u}_i$ 表示 OLS 残差。这就是人们熟悉的 F 检验统计量（*F*-**test stastistic**），它时常以残差平方和的形式重新表示。

在非线性模型中，甚至在线性模型中，要得出譬如 W_3 的准确结果是不可能的，因为它们需要非常强的假设。相反，利用 W_2 的非线性类似形式，并且其分布结果只是渐近形式。

7.2.2 非线性假设

考察 $q\times 1$ 维参数向量 $\boldsymbol{\theta}$ 的关于参数可能为非线性（**nonlinear in parameters**）的 h 个约束假设检验，其中，$h\leqslant q$。对线性回归来说，$\boldsymbol{\theta}=\boldsymbol{\beta}$且 $q=K$。

双侧零假设与备选假设分别是：

$$\begin{aligned} H_0&: \mathbf{h}(\boldsymbol{\theta})=\mathbf{0} \\ H_a&: \mathbf{h}(\boldsymbol{\theta})\neq\mathbf{0} \end{aligned} \tag{7.3}$$

其中，$\mathbf{h}(\cdot)$表示 $\boldsymbol{\theta}$ 的 $h\times 1$ 维向量函数。注意到，$\mathbf{h}(\boldsymbol{\theta})$在本章用于表示零假设约束。这里不应与前一章用于表示构建矩方法估计量或广义矩方法估计量的矩条件的 $\mathbf{h}(\mathbf{w},\boldsymbol{\theta})$相混淆。

熟悉的线性例子包括对单个系数 $h(\boldsymbol{\theta})=\theta_j=0$ 的统计显著性检验，以及对系数子集 $\mathbf{h}(\boldsymbol{\theta})=\boldsymbol{\theta}_2=\mathbf{0}$ 的检验。一个非线性的单个约束是 $h(\boldsymbol{\theta})=\theta_1/\theta_2-1=0$。后面几节将对这些例子加以研究。

假定 $h(\boldsymbol{\theta})$使得 $h\times q$ 阶矩阵：

$$\mathbf{R}(\boldsymbol{\theta})=\frac{\partial\mathbf{h}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'} \tag{7.4}$$

在 $\boldsymbol{\theta}=\boldsymbol{\theta}_0$ 处进行计算时，具有满秩 h，该假设等价于线性模型中约束的线性独立性，在这种情况下，$\mathbf{R}(\boldsymbol{\theta})=\mathbf{R}$ 并不依赖于 $\boldsymbol{\theta}$ 且具有满秩 h。假定在零假设下，参数没有位于参数空间的边界上（**boundary of the parameter space**）。这就是剔除了当模型需要 $\theta_1\geqslant 0$ 时对 H_0：$\theta_1=0$ 进行检验的情形。

7.2.3 沃尔德检验统计量

支持沃尔德检验的直觉极为简单。对 $\mathbf{h}(\boldsymbol{\theta}_0)=\mathbf{0}$ 是否成立的一个明显检验，是在没有利用约束时获得 $\hat{\boldsymbol{\theta}}$，然后查看 $\mathbf{h}(\hat{\boldsymbol{\theta}})\simeq 0$。若在 H_0 下，$h(\hat{\boldsymbol{\theta}})\overset{a}{\sim}\mathcal{N}[\mathbf{0},\mathrm{V}[\mathbf{h}(\hat{\boldsymbol{\theta}})]]$，

则检验统计量为：

$$\mathrm{W}=\mathbf{h}(\hat{\boldsymbol{\theta}})'[\mathrm{V}[\mathbf{h}(\hat{\boldsymbol{\theta}})]]^{-1}\mathbf{h}(\hat{\boldsymbol{\theta}})\overset{a}{\sim}\chi^2(h)$$

这里的唯一困难是求 $\mathrm{V}[\mathbf{h}(\hat{\boldsymbol{\theta}})]$，它将依赖于约束 $\mathbf{h}(\cdot)$ 和估计量 $\hat{\boldsymbol{\theta}}$。

在零假设下，由一阶泰勒级数展开式（参见 7.2.4 节）可知，$\mathbf{h}(\hat{\boldsymbol{\theta}})$ 具有与 $\mathbf{R}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$ 相同的极限分布，其中，$\mathbf{R}(\boldsymbol{\theta})$ 由式(7.4)定义。于是，在 H_0 下，$\mathbf{h}(\hat{\boldsymbol{\theta}})$ 服从渐近正态分布，其均值为 0，方差矩阵为 $\mathbf{R}(\boldsymbol{\theta}_0)\mathrm{V}[\hat{\boldsymbol{\theta}}]\mathbf{R}(\boldsymbol{\theta}_0)'$。一致估计值是 $\hat{\mathbf{R}}N^{-1}\hat{\mathbf{C}}\hat{\mathbf{R}}'$，其中，$\hat{\mathbf{R}}=\mathbf{R}(\hat{\boldsymbol{\theta}})$，假定估计量 $\hat{\boldsymbol{\theta}}$ 是 $\sqrt{N}$ 一致的，满足：

$$\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\xrightarrow{d}\mathcal{N}[\mathbf{0},\mathbf{C}_0] \tag{7.5}$$

而 $\hat{\mathbf{C}}$ 表示 $\mathbf{C}_0$ 的任何一致估计值。

沃尔德检验的普遍形式

由前面讨论，得到沃尔德检验统计量(**Wald test statistic**)：

$$\mathrm{W}=N\hat{\mathbf{h}}'[\hat{\mathbf{R}}\hat{\mathbf{C}}\hat{\mathbf{R}}']^{-1}\hat{\mathbf{h}} \tag{7.6}$$

其中，$\hat{\mathbf{h}}=\mathbf{h}(\hat{\boldsymbol{\theta}})$ 且 $\hat{\mathbf{R}}=\partial\mathbf{h}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'|_{\hat{\theta}}$。一个等价表达式是 $\mathrm{W}=\hat{\mathbf{h}}'[\hat{\mathbf{R}}\hat{\mathrm{V}}[\hat{\boldsymbol{\theta}}]\hat{\mathbf{R}}']^{-1}\hat{\mathbf{h}}$，其中，$\hat{\mathrm{V}}[\hat{\boldsymbol{\theta}}]=N^{-1}\hat{\mathbf{C}}$ 表示 $\hat{\boldsymbol{\theta}}$ 的估计渐近方差。

在 H_0 下，检验统计量 W 渐近服从 $\chi^2(h)$ 分布。因此，当 $\mathrm{W}>\chi^2_\alpha(h)$ 时，在显著水平 α 下，拒绝对应于 H_a 的 H_0；否则，不能拒绝 H_0。等价地讲，当 p 值(***p*-value**)小于 α 时，即 p 值等于 $\Pr[\chi^2(h)>\mathrm{W}]$，在水平 α 就拒绝 H_0。

人们还能把沃尔德检验统计量处理成为 F 检验。沃尔德渐近 F 统计量(**Wald asymptotic *F*-statistic**)

$$\mathrm{F}=\mathrm{W}/h \tag{7.7}$$

渐近地服从 $F(h,N-q)$ 分布。这就得出与式(7.6)中当 $N\rightarrow\infty$ 时的 W 相同的 p 值，尽管在有限样本下，p 值将会不一样。对非线性模型来说，最普遍报告的是 W，虽然 F 也时常用于小样本中，期待它提供较好的近似。

对仅有单个约束的检验来说，沃尔德卡方检验的平方根是标准正态检验统计量。该结果允许对单侧假设进行检验，故它十分有用。具体地讲，对于纯量 $h(\boldsymbol{\theta})$，沃尔德 z 检验统计量(**Wald *z*-test statistic**)是：

$$\mathrm{W}_z=\frac{\hat{h}}{\sqrt{\hat{\mathbf{r}}N^{-1}\hat{\mathbf{C}}\hat{\mathbf{r}}'}} \tag{7.8}$$

其中，$\hat{h}=h(\hat{\boldsymbol{\theta}})$，而 $\hat{\mathbf{r}}=\partial h(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'|_{\hat{\theta}}$ 表示 $1\times k$ 维向量。结果(7.6)意味着，W_z 在 H_0 下服从渐近标准正态分布。等价地讲，W_z 渐近服从 t 分布，其自由度为 $(N-q)$，因为当 $N\rightarrow\infty$ 时，t 是趋于正态的。因此，W_z 也可作为沃尔德 t 检验统计量。

讨论

非线性情况下的沃尔德检验统计量(7.6)与线性模型情况下由式(7.2)给出的统计量 W_2 具有同样形式。源自零假设的估计偏差是 $\mathbf{h}(\hat{\boldsymbol{\theta}})$ 而不是 $(\mathbf{R}\hat{\boldsymbol{\beta}}-\mathbf{r})$。矩阵

$\mathbf{R}$ 由估计导数矩阵 $\hat{\mathbf{R}}$ 代替，而 $\mathbf{R}$ 是满秩的假设则由 $\mathbf{R}_0$ 是满秩的假设代替。最后，估计量的估计渐近方差是 $N^{-1}\hat{\mathbf{C}}$，而不是 $s^2(\mathbf{X}'\mathbf{X})^{-1}$。

$\mathbf{C}_0$ 的一致估计值存在一个范围(参见 5.5.2 节)，实际应用时会得到渐近等价的 W、F 或 W_z 的各种不同计算值。特别地，$\mathbf{C}_0$ 经常具有三明治形式 $\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}$，通过稳健估计值 $\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}$ 一致地得到估计。沃尔德检验的优点是，在相对弱分布的假设——比如潜在异方差误差条件——下，很容易强有力地确保有效统计推断。

对于双侧检验，W_z、W 或 F 愈大，则愈可能拒绝 H_0。进一步情况会发生，即 $\mathbf{h}(\hat{\boldsymbol{\theta}})$ 源自零假设值 $\mathbf{0}$；估计量 $\hat{\boldsymbol{\theta}}$ 越有效，$\hat{\mathbf{C}}$ 就会越小；当样本量越大，则 N^{-1} 越小。这是当样本量增大并且在不改变显著性水平 α 时进行检验的结果。原则上讲，当样本量增大时，就能减小 α。完全参数模型的此类不利结果将在 8.5 节阐述。

7.2.4 *沃尔德统计量推导*

对于位于 $\hat{\boldsymbol{\theta}}$ 和 $\boldsymbol{\theta}_0$ 之间的某一个 $\boldsymbol{\theta}^+$，在 $\boldsymbol{\theta}_0$ 附近实施准确的一阶泰勒级数展开，得到：

$$\mathbf{h}(\hat{\boldsymbol{\theta}})=\mathbf{h}(\boldsymbol{\theta}_0)+\left.\frac{\partial \mathbf{h}}{\partial \boldsymbol{\theta}'}\right|_{\theta^+}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$$

由此可得：

$$\sqrt{N}(\mathbf{h}(\hat{\boldsymbol{\theta}})-\mathbf{h}(\boldsymbol{\theta}_0))=\mathbf{R}(\boldsymbol{\theta}^+)\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$$

其中，$\mathbf{R}(\boldsymbol{\theta})$已由式(7.4)定义，进而得到：

$$\sqrt{N}(\mathbf{h}(\hat{\boldsymbol{\theta}})-\mathbf{h}(\boldsymbol{\theta}_0))\xrightarrow{d}\mathcal{N}[\mathbf{0},\mathbf{R}_0\mathbf{C}_0\mathbf{R}_0'] \tag{7.9}$$

当 $\mathbf{R}(\boldsymbol{\theta}^+)\xrightarrow{p}\mathbf{R}=\mathbf{R}(\boldsymbol{\theta}_0)$时，这里直接应用极限正态乘积法则(定理 A.7)，同时利用由式(7.5)给出的$\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$极限分布。

在零假设下，由于 $\mathbf{h}(\boldsymbol{\theta}_0)=\mathbf{0}$，所以式(7.9)得以简化，故在 H_0 下，有：

$$\sqrt{N}\mathbf{h}(\hat{\boldsymbol{\theta}})\xrightarrow{d}\mathcal{N}[\mathbf{0},\mathbf{R}_0\mathbf{C}_0\mathbf{R}_0'] \tag{7.10}$$

在理论上，人们能使用多元变量正态分布定义拒绝区域，但一种更简单的方法是变换成卡方分布。回顾 $\mathbf{z}\sim\mathcal{N}[\mathbf{0},\boldsymbol{\Omega}]$，$\boldsymbol{\Omega}$ 是满秩的，这蕴含 $\mathbf{z}'\boldsymbol{\Omega}^{-1}\mathbf{z}\sim\chi^2(\dim(\boldsymbol{\Omega}))$。从而，在 H_0 下，由式(7.10)得出：

$$N\mathbf{h}(\hat{\boldsymbol{\theta}})'[\mathbf{R}_0\mathbf{C}_0\mathbf{R}_0']^{-1}\mathbf{h}(\hat{\boldsymbol{\theta}})\xrightarrow{d}\chi^2(h)$$

由 $\mathbf{R}_0$ 与 $\mathbf{C}_0$ 均是满秩的假设，此表达式中的矩阵逆存在。当 $\mathbf{R}_0$ 与 $\mathbf{C}_0$ 均由其一致估计值代替后，就获得由式(7.6)定义的沃尔德统计量。

7.2.5 *沃尔德检验例子*

最普遍的检验是对一个或多个排除性约束进行检验。我们还提供对非线性假设进行检验的一个例子。

对排除性约束检验

考察排除性约束，即 $\boldsymbol{\theta}$ 的最后 h 个分量等于 0。进而，$\mathbf{h}(\boldsymbol{\theta})=\boldsymbol{\theta}_2=\mathbf{0}$，这里，把 $\boldsymbol{\theta}$ 分割成 $\boldsymbol{\theta}=(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$。由此可得：

$$\mathbf{R}(\boldsymbol{\theta}) = \frac{\partial \mathbf{h}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} = \begin{bmatrix} \frac{\partial \boldsymbol{\theta}_2}{\partial \boldsymbol{\theta}_1'} & \frac{\partial \boldsymbol{\theta}_2}{\partial \boldsymbol{\theta}_2'} \end{bmatrix} = [\mathbf{0} \quad \mathbf{I}_h]$$

其中，$\mathbf{0}$ 表示 $(q-h)\times q$ 阶零矩阵，$\mathbf{I}_h$ 表示 $h\times h$ 阶单位阵，从而有：

$$\mathbf{R}(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\mathbf{R}(\boldsymbol{\theta})' = [\mathbf{0} \quad \mathbf{I}_h]\begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix}\begin{bmatrix} \mathbf{0} \\ \mathbf{I}_h \end{bmatrix} = \mathbf{C}_{22}$$

因此，排除性约束(**exclusion restrictions**)的沃尔德检验统计是：

$$\mathrm{W} = \hat{\boldsymbol{\theta}}_2'[N^{-1}\hat{\mathbf{C}}_{22}]^{-1}\hat{\boldsymbol{\theta}}_2 \tag{7.11}$$

其中，$N^{-1}\hat{\mathbf{C}}_{22}=\hat{\mathrm{V}}[\hat{\boldsymbol{\theta}}_2]$，且在 H_0 下渐近服从 $\chi^2(h)$分布。

该检验统计量是关于线性回归模型对回归元子集进行检验的推广。倘若误差服从正态分布且可使用有关的 F 检验，则在这种情况下就可利用小样本结果。

统计显著性检验

单个系数的显著性检验(**tests of significance of a single coefficient**)是对 $\boldsymbol{\theta}$ 的第 j 个分量 θ_j 是否异于零进行检验。于是，$h(\boldsymbol{\theta})=\theta_j$，$\mathbf{r}(\boldsymbol{\theta})=\partial h/\partial\boldsymbol{\theta}'$ 表示除第 j 个元素为 1 之外其余都为 0 的向量，因此，式(7.8)被简化为：

$$\mathrm{W}_z = \frac{\hat{\theta}_j}{\mathrm{se}[\hat{\theta}_j]} \tag{7.12}$$

其中，$\mathrm{se}[\hat{\theta}_j]=\sqrt{N^{-1}\hat{c}_{jj}}$ 表示 $\hat{\theta}_j$ 的标准误差，而 $\hat{c}_{jj}$ 表示 $\hat{\mathbf{C}}$ 的第 j 个对角元素。

归因于在正态性条件下线性回归模型的结果，式(7.12)中的检验统计量也经常被称为“t 统计量”，但严格地讲，它是渐近“z 统计量”(**z-statistic**)。

对 H_0 的双侧检验(**two-sided test**)来说，H_0：$\theta_{j0}=0$ 对应于 H_a：$\theta_{j0}\neq 0$，当 $|\mathrm{W}_z|>z_{\alpha/2}$，在显著性水平 α 上拒绝 H_0，否则就不能拒绝 H_0。这就得出与沃尔德卡方检验完全一致的结果，这是因为 $\mathrm{W}_z^2=\mathrm{W}$，其中，W 已由式(7.6)定义，且 $z_{\alpha/2}^2=\chi_\alpha^2(1)$。

关于 θ_j 符号，经常存在先验信息。因此，应使用单侧假设检验(**one-sided hypothesis test**)。例如，假定 $\theta_j>0$ 被认为是建立在经济推理或者过去研究的基础上。设定 $\theta_j>0$ 是零假设还是备选假设，这是有差异的。对于单侧检验来说，一种习惯做法是对所做出的判断设定为备选假设，因为可以证明，支持该判断需要较强的证据。在显著性水平 α 上，当 $\mathrm{W}_z>z_\alpha$，就拒绝 H_a：$\theta_{j0}>0$ 对应于 H_0：$\theta_{j0}\leqslant 0$。类似地，判断 $\theta_j<0$ 时，在显著性水平 α 上，对 H_0：$\theta_{j0}\geqslant 0$ 对应于 H_a：$\theta_{j0}<0$ 进行检验，当 $W_z<-z_\alpha$ 时，就拒绝 H_0。

对双侧检验来说，通常计算机会输出 p 值，但是在许多情况下，一种更合适的方法是使用单侧检验。若 $\hat{\theta}_j$ 具有“正确”的符号，则单侧检验 p 值就是所报告的双侧检验值的一半。

非线性约束检验

考察单个非线性约束的检验：

$$H_0: h(\boldsymbol{\theta})=\theta_1/\theta_2-1=0$$

于是，$\mathbf{R}(\boldsymbol{\theta})$表示 $1\times q$ 维向量，其第一元素为$\partial h/\partial\theta_1=1/\theta_2$，第二个元素为$\partial h/\partial\theta_2=-\theta_1/\theta_2^2$，而其余元素均为零。通过设$\hat{c}_{jk}$表示 $\hat{\mathbf{C}}$ 的第 jk 个元素，那么式(7.6)变成：

$$W=N\left(\frac{\hat{\theta}_1}{\hat{\theta}_2}\quad -1\right)^2\left(\begin{bmatrix}\frac{1}{\hat{\theta}_2} & -\frac{\hat{\theta}_1}{\hat{\theta}_2^2} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\hat{c}_{11} & \hat{c}_{12} & \cdots\\ \hat{c}_{21} & \hat{c}_{22} & \cdots\\ \vdots & \vdots & \ddots\end{bmatrix}\begin{bmatrix}1/\hat{\theta}_2\\ -\hat{\theta}_1/\hat{\theta}_2^2\\ \mathbf{0}\end{bmatrix}\right)^{-1}$$

其中，$\mathbf{0}$ 表示$(q-2)\times q$ 阶零矩阵，从而得到：

$$W=N[\hat{\theta}_2(\hat{\theta}_1-\hat{\theta}_2)]^2(\hat{\theta}_2^2\hat{c}_{11}-2\hat{\theta}_1\hat{\theta}_2\hat{c}_{12}+\hat{\theta}_1^2\hat{c}_{22})^{-1} \tag{7.13}$$

在 H_0 下，W 渐近服从 $\chi^2(1)$分布。等价地讲，$\sqrt{W}$渐近服从标准正态分布。

7.2.6 *错误设定模型的检验*

假设检验的大部分研究内容，包括本书第 7 章和第 8 章给出的内容，都假定除相对极少的并不会影响估计量一致性的错误设定且稳健的标准误差之外，零假设模型均是正确设定的。

实际上，这是过分简化的情况。例如，在对异方差误差进行检验时，假定这是回归不充分的唯一情况。可是，当条件均值被错误设定，检验的真实水平将不同于名义水平，甚至也不同于渐近情况。另外，检验的渐近等价性，诸如沃尔德检验、似然比检验以及拉格朗日乘子检验，都将不再成立。不过，模型设定得越好，检验就越有用。

同样地，注意到一些检验经常具有对应于假设不是以显形方式表述的备选假设的某种势。例如，假定零假设模型是 $y=\beta_1+\beta_2x+u$ ，其中，u 是同方差的。是否还包括 z 作为回归元的一个检验，也同样具有对应于模型关于 x 为非线性的备选假设的势，譬如 $y=\beta_1+\beta_2x+\beta_3x^2+u$，当 x 与 z 相关时。类似地，对应于异方差的检验，将同样具有对应关于 x_1 为非线性的某种势。对零假设拒绝，并不意味着备选假设模型是唯一的可行模型。

7.2.7 *联合检验与单独检验*

在应用研究中，人们经常想要知道源自系数集合的哪一个系数是“显著的”。当检验存在几种假设时，人们不是执行联合检验(**joint test**)或执行关注系数所有假设的联立检验，就是实施 F 检验假设的单独检验(**separate test**)。

线性回归的一个重要例子涉及，通过利用联合假设 H_0: $\beta_1=\beta_2=0$ 的 F 检验，对零假设 H_{10}: $\beta_1=0$ 与 H_{20}: $\beta_2=0$ 进行单独 t 检验的使用，其中的备选假设自始至终是，至少有一个参数不等于 0。倘若估计点$(\hat{\beta}_1, \hat{\beta}_2)$落在椭圆概率等高线外，则对拒绝 H_0 来说，F 检验是显性联合检验。或者，执行两个独立 t 检验。这种方法是隐性联合检验，称为诱导检验(**induced test**)[萨文(Savin，1984)]。如果不是

拒绝 H_{10}就是拒绝 H_{20},那么独立检验就要拒绝 H_0,若$(\hat{\beta}_1,\hat{\beta}_2)$落在其边界由两个检验统计量的临界值所构成的矩形之外,就会产生这种情况。尽管同样的显著性水平用于检验 H_0,因而椭圆与矩形具有相同的面积,但对联合检验与独立检验来说,其拒绝域则是不同的,从而它们之间存在潜在的矛盾。例如,$(\hat{\beta}_1,\hat{\beta}_2)$可能位于椭圆之内,而在矩形之外。

设 e_1 与 e_2 表示两个独立检验的第Ⅰ类错误的事件,并设 $e_I=e_1\cup e_2$ 表示诱导联合检验的第Ⅰ类错误的事件。从而,$\Pr[e_I]=\Pr[e_1]+\Pr[e_2]-\Pr[e_1\cap e_2]$,这蕴含:

$$\alpha_I\leqslant\alpha_1+\alpha_2 \tag{7.14}$$

其中,α_I、α_1 和 α_2 分别表示诱导联合检验、第一个独立检验以及第二个独立检验。在独立检验是统计独立的特别情况下,$\Pr[e_1\cap e_2]=\Pr[e_1]\Pr[e_2]=\alpha_1\alpha_2$,从而 $\alpha_I=\alpha_1+\alpha_2-\alpha_1\alpha_2$,对于 α_1 与 α_2 的典型小的值来说,譬如 0.05 或 0.01,$\alpha_1\alpha_2$ 是非常小的,而且上界(7.14)是检验水平的一个好标示变量。

相当多的诱导检验方面的文献都考察对独立检验临界值进行选取的问题,以使诱导检验具有已知水平。此处不长篇研讨这个问题,而是将邦费尼 t 检验(**Bonferroni *t*-test**)作为一个例子。该检验的临界值已被制成表格,参见萨文(Savin,1984)。

若信息矩阵的有关部分是对角的,则在具有正交回归元的线性回归中以及基于似然的检验(参见 7.3 节)中都会产生统计独立检验。那么,诱导联合检验统计量是建立在两个统计独立的独立检验统计量的基础上,而显性联合零检验统计量则是两个独立检验统计量之和。由于零检验的一个或者两个分量被拒绝,所以联合零检验可能被拒绝。运用独立检验将会揭示哪一种情况可以应用。

在更一般的相关回归元或者非对角信息矩阵情况下,显性联合检验具有拒绝零检验并且不能表明拒绝来源的缺点。假如运用诱导联合检验,对检验水平的设置就会需要邦费尼检验的某种变形或利用式(7.14)上界的近似。对每一个阶段都以前一阶段结果为条件,当顺次应用独立检验时,也会出现类似问题。8.7.1 节将阐述一种对具有两个假设的联合检验进行讨论的例子,那里检验的两个分量是相关的。

7.2.8 *置信区间方法*

用于推导沃尔德检验统计量的方法称为德尔塔方法(**delta method**,又称 δ 方法),这是因为$\mathbf{h}(\hat{\boldsymbol{\theta}})$的泰勒序列需要对 $\mathbf{h}(\boldsymbol{\theta})$求导数。该方法还能用于获得参数的非线性组合的分布,从而建立置信区间或区域。

第一个例子是通过 $\hat{\theta}_1/\hat{\theta}_2$ 估计比值 θ_1/θ_2。第二个例子是对条件均值 $g(\mathbf{x}'\boldsymbol{\beta})$进行预测,比如说,利用 $g(\mathbf{x}'\hat{\boldsymbol{\beta}})$预测 $g(\mathbf{x}'\boldsymbol{\beta})$。第三个例子是对 x 的一个分量变化进行弹性估计。

置信区间

参数向量 $\boldsymbol{\gamma}=\mathbf{h}(\boldsymbol{\theta})$的置信区间是通过

$$\hat{\gamma}=\mathbf{h}(\hat{\boldsymbol{\theta}}) \tag{7.15}$$

估计出的，其中，$\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$的极限分布已由式(7.5)给出。于是，直接应用式(7.9)可得到，$\sqrt{N}(\hat{\gamma}-\gamma_0)\xrightarrow{d}\mathcal{N}[\mathbf{0},\mathbf{R}_0\mathbf{C}_0\mathbf{R}_0']$，其中，$\mathbf{R}(\boldsymbol{\theta})$已由式(7.4)定义。等价地，称$\hat{\gamma}$渐近服从正态分布，其估计渐近方差矩阵为：

$$\mathrm{V}[\hat{\gamma}]=\hat{\mathbf{R}}N^{-1}\hat{\mathbf{C}}\hat{\mathbf{R}}' \tag{7.16}$$

此结果能用于建立置信区间或区域。

特别地，纯量参数γ的$100(1-\alpha)\%$置信区间(**confidence interval for the scalar parameter**)是：

$$\gamma\in\hat{\gamma}\pm z_{\alpha/2}se[\hat{\gamma}] \tag{7.17}$$

其中：

$$\mathrm{se}[\hat{\gamma}]=\sqrt{\hat{\mathbf{r}}N^{-1}\hat{\mathbf{C}}\hat{\mathbf{r}}'} \tag{7.18}$$

这里，$\hat{\mathbf{r}}=\mathbf{r}(\hat{\boldsymbol{\theta}})$，而$\mathbf{r}(\boldsymbol{\theta})=\partial\gamma/\partial\boldsymbol{\theta}'=\partial h(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'$。

置信区间例子

举一个例子，假定$\mathrm{E}[y|\mathbf{x}]=\exp(\mathbf{x}'\boldsymbol{\beta})$，并想要获得当$\mathbf{x}=\mathbf{x}_p$时预测条件均值的置信区间。于是，$h(\boldsymbol{\beta})=\exp(\mathbf{x}'\boldsymbol{\beta})$，因而$\partial h/\partial\boldsymbol{\beta}'=\exp(\mathbf{x}_p'\boldsymbol{\beta})\mathbf{x}_p$，且由式(7.18)得到：

$$\mathrm{se}[\exp(\mathbf{x}_p'\hat{\boldsymbol{\beta}})]=\exp(\mathbf{x}_p'\hat{\boldsymbol{\beta}})\sqrt{\mathbf{x}_p'N^{-1}\hat{\mathbf{C}}\mathbf{x}_p}$$

其中，$\hat{\mathbf{C}}$表示$\sqrt{N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)$的极限分布中方差矩阵的一致估计值。

举第二个例子，假定想要得到e^{β}的置信区间，而不是纯量系数β的置信区间。那么，$h(\beta)=e^{\beta}$，因而$\partial h/\partial\beta=e^{\beta}$，并由式(7.18)得到，$\mathrm{se}[e^{\hat{\beta}}]=e^{\hat{\beta}}\mathrm{se}[\hat{\beta}]$。这就得出$e^{\beta}$的95%置信区间在$e^{\hat{\beta}}\pm1.96e^{\hat{\beta}}\mathrm{se}[\hat{\beta}]$之间。

δ方法并不总是获得置信区间的最佳方法，这是因为它把置信区间限制成关于$\hat{\gamma}$对称的。此外，在前面例子中，尽管$e^{\beta}>0$，但其置信区间可通过对β置信区间中的项取指数而获得。从而有：

$$\Pr[\hat{\beta}-1.96\mathrm{se}[\hat{\beta}]<\beta<\hat{\beta}+1.96\mathrm{se}[\hat{\beta}]]=0.95$$
$$\Rightarrow\Pr[\exp(\hat{\beta}-1.96\mathrm{se}[\hat{\beta}])<e^{\beta}<\exp(\hat{\beta}+1.96\mathrm{se}[\hat{\beta}])]=0.95$$

该置信区间具有非对称且仅包括正值的优点。这种变换经常用于二值模型以及持续期间模型的斜率参数的置信区间。当$h(\cdot)$是单调的时，将这一方法推广到其他变换$\gamma=h(\theta)$。

7.2.9 *沃尔德检验缺少不变性*

若能获得无约束模型的估计值，就容易得出沃尔德检验统计量，而且该统计量的效力不亚于其他可行的检验方法，如同稍后章节所讨论的。因此，它是最普遍使用的检验方法。

可是，沃尔德检验存在一个基本问题：对零假设在代数形式上等价的参数化来说，它不是不变的。例如，考察7.2.5节的例子。于是，H_0：$\theta_1/\theta_2-1=0$能等价地

表述成 $H_0: \theta_1-\theta_2=0$，得到沃尔德卡方检验统计量：

$$W^*=N(\hat{\theta}_1-\hat{\theta}_2)^2(\hat{c}_{11}-2\hat{c}_{12}+\hat{c}_{22})^{-1} \quad (7.19)$$

它不同于式(7.13)中的 W。尽管 W 与 W^* 是渐近等价的，但有限样本中的统计量 W 与 W^* 相差甚远。如同格雷戈里和维尔(Gregory and Veall, 1985)曾经考察的一个非常类似的例子，他们运用蒙特卡罗模拟研究表明，小样本差异性是相当大的。对名义水平为 0.05 的检验来说，沃尔德检验的一种变形通过所有模拟都具有 0.04 与 0.06 之间的实际水平，因而渐近理论提供了一种小样本的良好近似，而一种可供选择的沃尔德检验的渐近等价变形在某些模拟中具有超过 0.20 的实际水平。

菲利普斯和帕克(Phillips and Park, 1988)解释了这一差异，他们表明，利用传统渐近方法，尽管各种不同零假设约束表示具有相同的卡方分布，可是一旦利用更精炼的建立在埃奇沃思展开式基础上的渐近理论(参见 11.4.3 节)，它们却服从不同的渐近分布。进一步地，在特殊背景下，诸如前面例子，埃奇沃思展开式能用于表明 H_0 的参数化以及参数空间的区域，通常渐近理论可能提供一个不好的小样本近似。

当对非线性约束进行检验时，其经验教训是要小心谨慎。作为一种稳健性检查，人们能利用各种不同的在代数形式上等价的零假设约束表示，来执行几种沃尔德检验。若这些检验得出实质上截然不同的结论，就可能存在问题。一种解决方法是执行沃尔德检验的自助法形式。这就提供了较好的小样本特性，同时剔除利用各种不同的 H_0 表示的沃尔德检验之间的大部分差异，因为由 11.4.4 节知，自助法本质上执行埃奇沃思展开式。第二种解决方法是运用下一节给出的其他检验方法，这些方法对于 H_0 的各种不同表示是不变的。

7.3 基于似然的检验

本节考察，已知似然函数——其分布是完全设定的——时的假设检验问题。于是，存在三种实施假设检验的经典统计方法：沃尔德检验、似然比(LR)检验以及拉格朗日乘子(LM)检验。第四种检验是归功于内曼(Neyman, 1959)的 $\mathbf{C}(\alpha)$ 检验，该检验并不普遍使用，所以这里没有阐述；参见戴维森和麦金农(Davidson and McKinnon, 1993)。所有这四种检验均是渐近等价的，因而对它们进行选取，要考虑到计算的方便与否以及有限样本特性。本节还没有涵盖内曼(Neyman, 1937)的光滑检验(**smooth test**)，贝拉和戈什(Brea and Ghosh, 2002)曾讨论哪一种是最优的且像其他检验一样是基本的。

这些结果均假定似然函数被正确设定。7.5 节给出对建立在拟极大似然估计量、m 估计量以及有效广义矩方法估计量基础上检验的推广。

7.3.1 *沃尔德检验、似然比检验以及拉格朗日（得分）检验*

设 $L(\boldsymbol{\theta})$ 表示似然函数，即给定 $\mathbf{X}$ 与 $\boldsymbol{\theta}$ 参数时 $\mathbf{y}$ 的联合条件密度。我们想要检验式(7.3)给出的零假设：$\mathbf{h}(\boldsymbol{\theta}_0)=\mathbf{0}$。

除沃尔德检验之外，其他检验都要求利用零假设约束进行估计。定义估计量：

$$\begin{aligned}&\hat{\boldsymbol{\theta}}_u \quad (\text{无约束 MLE})\\&\tilde{\boldsymbol{\theta}}_r \quad (\text{约束 MLE})\end{aligned} \tag{7.20}$$

无约束极大似然估计量(**unrestricted MLE**)$\hat{\boldsymbol{\theta}}_u$ 是对 $\ln L(\boldsymbol{\theta})$ 求极大值；在对沃尔德检验的线性讨论时，更简单地记为 $\hat{\boldsymbol{\theta}}$ 。约束极大似然估计量 $\tilde{\boldsymbol{\theta}}_r$ 是对拉格朗日算子 $\ln L(\boldsymbol{\theta})-\boldsymbol{\lambda}'\mathbf{h}(\boldsymbol{\theta})$ 求极大值，其中，$\boldsymbol{\lambda}$ 表示 $h\times1$ 维拉格朗日乘子。在排除性约束 $\mathbf{h}(\boldsymbol{\theta})=\boldsymbol{\theta}_2=\mathbf{0}$ 的简单情况下，其中，$\boldsymbol{\theta}=(\boldsymbol{\theta}_1',\boldsymbol{\theta}_2')$，约束极大似然估计量是 $\tilde{\boldsymbol{\theta}}_r=(\tilde{\boldsymbol{\theta}}_{1r}',\mathbf{0}')$，这里，$\tilde{\boldsymbol{\theta}}_{1r}'$是对约束似然 $\ln L(\boldsymbol{\theta}_1,\mathbf{0})$求关于 $\boldsymbol{\theta}_1$ 的极大值而直接获得的，而 $\mathbf{0}$ 表示 $(q-h)\times1$ 维零向量。

这里将引出并定义三种检验统计量，有关其推导则推迟到 7.3.3 节。所有这三种统计量在 H_0 下均依分布收敛到 $\chi^2(h)$ 。因此，当计算的检验统计量大于 $\chi^2_\alpha(h)$，在显著性水平 α 上拒绝 H_0。等价地讲，当 $p\leqslant\alpha$ 时，在水平 α 上拒绝 H_0，其中，$p=\Pr[\chi^2(h)>t]$表示 p 值，而 t 表示检验统计量的计算值。

似然比检验

激发似然比检验统计量的是，当 H_0 正确时，无约束的对数似然函数的极大值与约束的对数似然函数的极大值是同样的。这建议利用 $\ln L(\hat{\boldsymbol{\theta}}_u)$与 $\ln L(\tilde{\boldsymbol{\theta}}_r)$之差的函数。

实施检验需要获得该差的极限分布。可以证明，在 H_0 下，2 倍差分服从渐近卡方分布。从而，立刻得出似然比检验(**likelihood ration test**)统计量：

$$\mathrm{LR}=-2[\ln L(\tilde{\boldsymbol{\theta}}_r)-\ln L(\hat{\boldsymbol{\theta}}_u)] \tag{7.21}$$

沃尔德检验

激发沃尔德检验的动机是，当 H_0 正确时，无约束极大似然估计量 $\hat{\boldsymbol{\theta}}_u$ 应满足 H_0 约束，因此，$h(\hat{\boldsymbol{\theta}}_u)$应接近于 0。

实施检验需要获得 $\mathbf{h}(\hat{\boldsymbol{\theta}}_u)$的渐近分布。沃尔德检验的一般形式已由式(7.6)给出。对极大似然估计而言，通过信息矩阵(IM)等式 $\mathrm{V}[\hat{\boldsymbol{\theta}}_u]=-N^{-1}\mathbf{A}_0^{-1}$ 会得出特殊化结果，其中：

$$\mathbf{A}_0=\operatorname{plim} N^{-1}\left.\frac{\partial^2\ln L}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\theta_0} \tag{7.22}$$

这就得出沃尔德检验(**Wald test**)统计量：

$$\mathrm{W}=-N\hat{\mathbf{h}}'[\hat{\mathbf{R}}\hat{\mathbf{A}}^{-1}\hat{\mathbf{R}}']^{-1}\hat{\mathbf{h}} \tag{7.23}$$

其中，$\hat{\mathbf{h}}=\mathbf{h}(\hat{\boldsymbol{\theta}}_u)$，$\hat{\mathbf{R}}=\mathbf{R}(\hat{\boldsymbol{\theta}}_u)$，$\mathbf{R}(\boldsymbol{\theta})=\partial h(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'$，而 $\hat{\mathbf{A}}$ 表示 $\mathbf{A}_0$ 的一致估计值。由于 $\mathbf{A}_0$ 是正定的，故出现负号。

拉格朗日乘子检验或得分检验

引出拉格朗日乘子检验统计量的动机是，在似然函数极大值处的梯度 $\partial\ln L/\partial\boldsymbol{\theta}|_{\hat{\theta}_u}=\mathbf{0}$。当 H_0 正确时，这个极大值也应在约束极大似然估计处(即 $\partial\ln L/\partial\boldsymbol{\theta}|_{\tilde{\theta}_u}\simeq\mathbf{0}$)达到，因此，利用此约束极少对 $\boldsymbol{\theta}$ 估计值产生影响。源于此动机，

由于$\partial \ln \theta/\partial \boldsymbol{\theta}$是得分向量，故拉格朗日乘子被称为得分检验[1]（**score test**）。

一种可供选择的动机是测量关于约束极大似然估计的约束最优化问题的拉格朗日乘子（**Lagrange multipliers**）接近于0的情况。对$\ln L(\boldsymbol{\theta})-\boldsymbol{\lambda}'\mathbf{h}(\boldsymbol{\theta})$求关于$\boldsymbol{\theta}$的极大值，得出：

$$\left.\frac{\partial \ln L}{\partial \boldsymbol{\theta}}\right|_{\tilde{\theta}_r}=\left.\frac{\partial \mathbf{h}(\boldsymbol{\theta})'}{\partial \boldsymbol{\theta}}\right|_{\tilde{\theta}_r}\times\tilde{\boldsymbol{\lambda}}_r \tag{7.24}$$

由此可得，建立在估计拉格朗日乘子$\tilde{\boldsymbol{\lambda}}_r$基础上的检验，等价于建立在$\partial \ln L/\partial \boldsymbol{\theta}|_{\tilde{\theta}_r}$基础上的检验，这是因为，假定$\partial \mathbf{h}/\partial \boldsymbol{\theta}'$是满秩的。

实施检验需要获得$\partial \ln L/\partial \boldsymbol{\theta}|_{\tilde{\theta}_r}$的渐近分布。这就得出拉格朗日乘子检验（**Lagrange multiplier test**）或得分检验（**score test**）统计量：

$$\mathrm{LM}=-N^{-1}\left.\frac{\partial \ln L}{\partial \boldsymbol{\theta}'}\right|_{\tilde{\theta}_r}\tilde{\mathbf{A}}^{-1}\left.\frac{\partial \ln L}{\partial \boldsymbol{\theta}}\right|_{\tilde{\theta}_r} \tag{7.25}$$

其中，$\tilde{\mathbf{A}}$表示式(7.22)在$\mathbf{A}_0$处而不是在$\tilde{\boldsymbol{\theta}}_u$处计算的一致估计值。

归功于艾奇逊和西尔维（Aitchison and Silvey，1958）与西尔维（Silvey，1959）的LM检验（**LM test**），等价于拉奥（Rao，1947）的得分检验。检验统计量LM通常是通过获得得分的解析表达式而不是拉格朗日乘子推导出的。尽管比较清晰的术语为得分检验，但经济计量学家通常还是称该检验为LM检验。

讨论

布斯（Buse，1982）通过对这三种检验进行图示阐述，提供了一种很好的直观效果，他把全部三种检验处理成对数似然变化的测量。这里，我们提供一种语言描述性概述。

考察纯量参数以及$\theta_0-\theta^*=0$是否成立的沃尔德检验。于是，$\hat{\beta}_u$与θ^*的已知不同将被转化成$\ln L$上的较大变化，具有较大曲度的就是对数似然函数。曲率的一种正常测度是二阶导数$H(\theta)=\partial^2 \ln L/\partial\theta^2$。这就建议$\mathrm{W}=-(\hat{\theta}_u-\theta^*)^2H(\hat{\theta}_u)$。式(7.23)中的统计量W能被看成对向量$\boldsymbol{\theta}$与带有测量曲率$N\hat{\mathbf{A}}$的更一般约束$\mathbf{h}(\boldsymbol{\theta}_0)$的推广。

对得分检验来说，布斯已经证明，$\partial \ln L/\partial\theta|_{\tilde{\theta}_r}$的已知值被转换成$\ln L$的较大变化，具有较小曲度的就是对数似然函数。这就导致在式(7.25)中运用了$(N\tilde{\mathbf{A}})^{-1}$。而且，统计量LR可直接与对数似然进行比较。

例子

为了阐明这三种检验，考察一个满足$y_i\sim\mathcal{N}[\mu_0,1]$且检验$H_0$：$\mu_0=\mu^*$的iid例子。于是，$\hat{\mu}_u=\bar{y}$且$\tilde{\mu}_r=\mu^*$。

对于拉格朗日乘子检验，$\ln L(\mu)=-\frac{N}{2}\ln 2\pi-\frac{1}{2}\sum_i(y_i-\mu)^2$，经过一些代数运算，得到：

$$\mathrm{LR}=2[\ln L(\bar{y})-\ln L(\mu^*)]=N(\bar{y}-\mu^*)^2$$

---

[1] 又称为分值检验。——译者注

沃尔德检验是建立在 $\bar{y}-\mu^* \simeq 0$ 是否成立的基础上。这里容易证明，在 H_0 下，$\bar{y}-\mu^* \sim \mathcal{N}[0,1/N]$，从而得到二次形式：

$$W=(\bar{y}-\mu^*)[1/N]^{-1}(\bar{y}-\mu^*)$$

该式被简化成 $N(\bar{y}-\mu^*)^2$，进而 W=LR。

拉格朗日乘子检验(LM)是建立在$\partial \ln L(\mu)/\partial\mu|_{\mu^*}=\sum_i(y_i-\mu_0)|_{\mu^*}=N(\bar{y}-\mu^*)$ 接近于 0 的基础上。这恰好是对$(\bar{y}-\mu^*)$的重新标度，所以 LM=W。更正式地讲，$\widetilde{A}(\mu^*)=-1$，由于$\partial^2 \ln L(\mu)/\partial\mu^2=-N$，且由式(7.25)得到：

$$LM=N^{-1}(N(\bar{y}-\mu^*))[1]^{-1}(N(\bar{y}-\mu^*))$$

它可被简化成 $N(\bar{y}-\mu^*)^2$，从而验证 LM=W=LR。

尽管这三种检验具有截然不同的产生动机，但三种检验统计量在此处却是相同的。归因于对数似然关于 μ 为二次的，这种精确等价是常值曲率的特殊例子。更一般地讲，三种检验统计量在有限样本中是不同的，但它们是渐近等价的(参见 7.3.4 节)。

7.3.2 *泊松回归例子*

考察 5.2 节引入的泊松回归模型中排除性约束的检验。这个例子主要是考虑到教学上的方便，因为实际上人们应该在与那些泊松模型相比为弱分布的假设下，实施计数数据的统计推断(参见第 20 章)。

若给定 $\mathbf{x}$ 时 y 服从泊松分布，其条件均值为 $\exp(\mathbf{x}'\boldsymbol{\beta})$，则对数似然函数是：

$$\ln L(\boldsymbol{\beta}) = \sum_{i=1}^{N}\{-\exp(\mathbf{x}_i'\boldsymbol{\beta})+y_i\mathbf{x}_i'\boldsymbol{\beta}-\ln y_i!\} \tag{7.26}$$

对 h 个排除性约束来说，其零假设是 $H_0: \mathbf{h}(\boldsymbol{\beta})=\boldsymbol{\beta}_2=\mathbf{0}$，其中，$\boldsymbol{\beta}=\boldsymbol{\beta}_1',\boldsymbol{\beta}_2'$)。

无约束 MLE $\hat{\boldsymbol{\beta}}$ 对式(7.26)求关于 $\boldsymbol{\beta}$ 极大值，并具有一阶条件 $\sum_i(y_i-\exp(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{x}_i=\mathbf{0}$。该极限方差矩阵是$-\mathbf{A}^{-1}$，其中：

$$\hat{\mathbf{A}}=-\operatorname{plim} N^{-1}\sum_i \exp(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i\mathbf{x}_i'$$

约束 MLE 是$\widetilde{\boldsymbol{\beta}}=(\widetilde{\boldsymbol{\beta}}_1',\mathbf{0}')'$，其中，$\widetilde{\boldsymbol{\beta}}_1$ 对式(7.26)求关于 $\boldsymbol{\beta}_1$ 的极大值，由于 $\boldsymbol{\beta}_2=\mathbf{0}$，故用 $\mathbf{x}_{1i}'\boldsymbol{\beta}_1$ 代替 $\mathbf{x}_i'\boldsymbol{\beta}$。因而，$\widetilde{\boldsymbol{\beta}}_1$ 是一阶条件$\sum_i(y_i-\exp(\mathbf{x}_{1i}'\boldsymbol{\beta}_1)\mathbf{x}_{1i}=\mathbf{0}$ 的解。

似然比检验统计量(7.21)很容易从约束模型的与无约束模型的拟合对数似然中计算出来。

源于 7.2.5 节的排除性约束的沃尔德检验统计量是 $W=-N\hat{\boldsymbol{\beta}}_2'\hat{\mathbf{A}}^{22}\hat{\boldsymbol{\beta}}_2$，其中，$\hat{\mathbf{A}}^{22}$表示 $\hat{\mathbf{A}}^{-1}$的(2, 2)分块，并且 $\hat{\mathbf{A}}=-N^{-1}\sum_i\exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}})\mathbf{x}_i\mathbf{x}_i'$。

LM 检验是建立在$\partial \ln L(\boldsymbol{\beta})/\partial\boldsymbol{\beta}=\sum_i\mathbf{x}_i(y_i-\exp(\mathbf{x}_i'\boldsymbol{\beta}))$的基础上。在约束 MLE 处，它等于$\sum_i\mathbf{x}_i\tilde{u}_i$，其中，$\tilde{u}_i=y_i-\exp(\mathbf{x}_{1i}'\widetilde{\boldsymbol{\beta}}_1)$ 表示源自对约束模型进行估计的残差。LM 检验统计量(7.25)是：

$$LM=\left[\sum_{i=1}^{N}\mathbf{x}_i\tilde{u}_i\right]'\left[\sum_{i=1}^{N}\exp(\mathbf{x}_{1i}'\widetilde{\boldsymbol{\beta}}_1)\mathbf{x}_i\mathbf{x}_i'\right]^{-1}\left[\sum_{i=1}^{N}\mathbf{x}_i\tilde{u}_i\right] \tag{7.27}$$

由前面给定的约束 MLE 的一阶条件知，由于$\sum_i \mathbf{x}_{1i}\tilde{u}_i=\mathbf{0}$，可能得到某种进一步简化。这里的 LM 检验建立在省略回归元与其残差相关的基础上，其结果可推广到 7.3.5 节的其他例子。

通常很难获得 LM 检验的代数表达式。对 LM 检验的标准应用来说，就是这样做的，并纳入计算机软件包之中。通过辅助回归，进行计算也是可能的(参见 7.3.5 节)。

7.3.3 *推导检验*

沃尔德检验的分布已在 7.2.4 节正式推导出来。有关似然比检验与拉格朗日检验的证明则更为复杂，此处我们仅仅概述其证明。

似然比检验

为了简单起见，考察零假设是 $\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}$ 的特殊情况，因此，$\tilde{\boldsymbol{\theta}}_r=\bar{\boldsymbol{\theta}}$ 时不存在任何估计误差。在 $\ln L(\bar{\boldsymbol{\theta}})$附近取 $\ln L(\hat{\boldsymbol{\theta}}_u)$ 的二阶泰勒级数展开式，得到：

$$\ln L(\bar{\boldsymbol{\theta}})=\ln L(\hat{\boldsymbol{\theta}}_u)+\left.\frac{\partial \ln L}{\partial \boldsymbol{\theta}'}\right|_{\hat{\theta}_u}(\bar{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}_u)+\frac{1}{2}(\bar{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}_u)'\left.\frac{\partial^2 \ln L}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right|_{\hat{\theta}_u}(\bar{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}_u)+R$$

其中，R 表示剩余项。由一阶条件知，$\partial \ln L/\partial \boldsymbol{\theta}|_{\hat{\theta}_u}=\mathbf{0}$，经过重新整理得出：

$$-2[\ln L(\bar{\boldsymbol{\theta}})-\ln L(\hat{\boldsymbol{\theta}}_u)]=-(\bar{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}_u)'\left.\frac{\partial^2 \ln L}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right|_{\hat{\theta}_u}(\bar{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}_u)+R \qquad (7.28)$$

在 H_0：$\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}$ 下，由标准结果知，$\sqrt{N}(\hat{\boldsymbol{\theta}}_u-\bar{\boldsymbol{\theta}})\xrightarrow{d}\mathcal{N}[\mathbf{0},-[\text{plim}\,N^{-1}\partial^2\ln L/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}']^{-1}]$，故式(7.28)右边服从 $\chi^2(h)$。例如，在一般情况下，关于 LR 极限分布的推导，参见雨宫(Amemiya，1985，第 143 页)。

偏爱 LR 的一个理由是，由内曼—皮尔逊(Neyman-Pearson，1933)引理可知，对简单零假设对应于简单备选假设进行检验的始终最有效力的就是似然比 $L(\tilde{\boldsymbol{\theta}}_r)/L(\hat{\boldsymbol{\theta}}_u)$函数，尽管特定函数$-2\ln(L(\tilde{\boldsymbol{\theta}}_r)/L(\hat{\boldsymbol{\theta}}_u))$不一定等于式(7.21)给出的似然比，而且可对该统计量给出一个称谓。

LM 或得分检验

由一阶泰勒级数展开式可知：

$$\frac{1}{\sqrt{N}}\left.\frac{\partial \ln L}{\partial \boldsymbol{\theta}}\right|_{\tilde{\theta}_r}=\frac{1}{\sqrt{N}}\left.\frac{\partial \ln L}{\partial \boldsymbol{\theta}}\right|_{\theta_0}+\frac{1}{N}\frac{\partial^2 \ln L}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\sqrt{N}(\tilde{\boldsymbol{\theta}}_r-\boldsymbol{\theta}_0)$$

而其右边的两项有助于得出极限分布。于是，可以证明，由式(7.25)定义的拉格朗日乘子的 $\chi^2(h)$分布遵从：

$$\mathbf{R}_0\mathbf{A}_0^{-1}\frac{1}{\sqrt{N}}\left.\frac{\partial \ln L}{\partial \boldsymbol{\theta}}\right|_{\tilde{\theta}_r}\xrightarrow{d}\mathcal{N}[\mathbf{0},\mathbf{R}_0\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}\mathbf{R}_0'] \qquad (7.29)$$

其详细推导已由伍德里奇(Wooldridge，2002，第 365 页)给出，例如，$\mathbf{R}_0$ 与 $\mathbf{A}_0$ 已经由式(7.4)与式(7.22)定义，而：

$$\mathbf{B}_0=\text{plim}\,N^{-1}\left.\frac{\partial \ln L}{\partial \boldsymbol{\theta}}\frac{\partial \ln L}{\partial \boldsymbol{\theta}'}\right|_{\theta_0} \qquad (7.30)$$

结果(7.29)得出比式(7.25)更复杂的卡方统计量，但通过信息矩阵等式可对它简化为式(7.25)。

7.3.4 *哪个检验好*

通常，对检验方法的选择要依据稳健形式存在与否、有限样本特性以及计算是否简单来决定。

渐近等价性

所有这三种检验统计量在 H_0 下都服从渐近 $\chi^2(h)$ 分布。进一步地，可以证明，三种统计量都是非中心 $\chi^2(h;\lambda)$ 分布，并且在局部备选条件下，具有相同的非中心性参数。7.6.3 节给出沃尔德检验的详细内容。因此，这三种检验对应于局部备选假设具有相同渐近势。

三种统计量的有限样本分布并不一样。在具有正态性的线性回归模型中，对于 $\boldsymbol{\theta}$ 的 h 个线性约束的沃尔德检验的一种变形，等于 $F(h,N-K)$ 统计量(参见 7.2.1 节)，可是似然比统计量与拉格朗日乘子统计量却都不存在解析结果。更一般地讲，非线性模型不存在准确的小样本结果。

在一些情况下，能得出这三种统计量取值的次序。尤其是在正态性下，就线性回归模型的线性约束检验而言，波恩特和萨文(Berndt and Savin，1977)已经证明，沃尔德检验≥LR≥LM。这个结果很少具有理论重要性，因为在零假设下，最不可能拒绝的检验将具有最小的实际水平，但也具有最小势。不过，它对于线性模型却具有实践重要性，因为它意味着当在固定名义水平上进行检验时，沃尔德检验将总是比似然比更经常地拒绝 H_0，同样，似然比检验将总是比拉格朗日乘子检验更经常地拒绝 H_0。研究者更偏爱用沃尔德检验来决定拒绝 H_0。此结果局限于线性模型。

重新参数化的不变性

对零假设的代数形式上等价的重新参数化来说，沃尔德检验不是不变的(参见 7.2.9 节)，而 LR 检验是不变的。可是，不是所有的拉格朗日乘子检验形式都是不变的。当期望海赛矩阵(参见 5.5.2 节)用于估计 $\mathbf{A}_0$，拉格朗日乘子检验通常是不变的；而当海赛矩阵用于估计 $\mathbf{A}_0$，拉格朗日乘子检验就不是不变的。稍后由式(7.34)定义的 LM^* 检验是不变的。沃尔德检验缺乏不变性是其主要弱点。

稳健形式

在错误设定密度的一些情况下，准 MLE(参见 5.7 节)仍是一致的。于是，LM 很容易得以稳健处理(参见 7.2 节)。对拉格朗日乘子检验就更难进行稳健处理，参见 7.5.1 节的 m 估计量一般结果以及 8.4 节关于稳健拉格朗日乘子检验的一些例子。除了在稍后由式(7.39)给出的特殊情况下，似然比检验不再服从卡方分布。相反，拉格朗日乘子检验却服从卡方分布的混合形式(参见 8.5.3 节)。

简便

计算简便也是一个要考虑的因素。似然比需要对模型计算两次，其中一次是对无约束的零假设，另一次则是对约束的零假设。当运用软件计算时就很容易，因为人们只需要读出例行打印出的对数似然、相减并乘 2 就可以。当无约束模型容

易估计时，沃尔德检验只需要在 H_a 下进行估计，而且是最佳的。例如，这正是对非线性模型中条件均值参数限制的情况，譬如 NLS、probit、Tobit 以及 logit。当约束模型容易估计时，拉格朗日乘子统计量只需在 H_0 下进行估计，而且是最佳的。一些例子就是对自相关与异方差性进行检验，其最容易的是估计不具有这些复杂情况的零假设模型。

沃尔德检验经常用来对统计显著性进行检验，而拉格朗日乘子检验则时常用来对正确设定模型进行检验。

7.3.5 *拉格朗日乘子检验的解释与计算*

在一些重要例子中，拉格朗日乘子检验具有额外的简单解释以及通过辅助回归进行计算的优点。

本节关注的内容限制在针对不同 i 的独立的通常的横截面数据情况的纯量因变量上，因而有 $\ln L(\boldsymbol{\theta})/\partial\boldsymbol{\theta}=\sum_i s_i(\boldsymbol{\theta})$，其中：

$$\mathbf{s}_i(\boldsymbol{\theta})=\frac{\partial\ln f(y_i|\mathbf{x}_i,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \tag{7.31}$$

表示第 i 个观测值对无约束模型得分向量的贡献。由式(7.25)知，LM 检验就是对 $\sum_i s_i(\tilde{\boldsymbol{\theta}}_r)$ 接近于 0 进行检验。

拉格朗日乘子检验的简单解释

假定密度使得 $\mathbf{s}(\boldsymbol{\theta})$ 因式分解为：

$$\mathbf{s}(\boldsymbol{\theta})=\mathbf{g}(\mathbf{x},\boldsymbol{\theta})r(y,\mathbf{x},\boldsymbol{\theta}) \tag{7.32}$$

这里，对于某个 $q\times1$ 维向量函数 $\mathbf{g}(\cdot)$ 以及纯量函数 $r(y,\mathbf{x},\boldsymbol{\theta})$，因为 y 出现在 $r(\cdot)$ 中而未出现在 $\mathbf{g}(\cdot)$ 中，故 $r(y,\mathbf{x},\boldsymbol{\theta})$ 可被解释成广义残差。例如，对泊松回归来说，$\partial\ln f/\partial\boldsymbol{\theta}=\mathbf{x}(y-\exp(\mathbf{x}'\boldsymbol{\beta}))$。

已知式(7.32)以及对不同 i 的独立性，则有 $\partial\ln L/\partial\boldsymbol{\theta}\mid_{\tilde{\theta}_r}=\sum_i\tilde{\mathbf{g}}_i\tilde{r}_i$，其中，$\tilde{\mathbf{g}}_i=\mathbf{g}(\mathbf{x},\tilde{\boldsymbol{\theta}}_r)$ 而 $\tilde{r}_i=r(y_i,\mathbf{x}_i,\tilde{\boldsymbol{\theta}}_r)$。因此，拉格朗日乘子检验可简单地被解释为，对 $\tilde{\mathbf{g}}_i$ 与残差 $\tilde{r}_i$ 之间相关性的得分检验。在 7.3.2 节带有泊松回归的拉格朗日乘子中，已给出这种解释，其中，$\tilde{\mathbf{g}}_i=\mathbf{x}_i$ 且 $\tilde{r}_i=y_i-\exp(\mathbf{x}_{1i}'\tilde{\boldsymbol{\beta}}_1)$。

每当 $f(y)$ 建立在一个参数密度的基础上，就会得到分解(7.32)。尤其是，许多普遍的似然模型均是建立在一个参数 LEF 密度上，其参数为 μ，从而建模成 $\mathbf{x}$ 与 $\boldsymbol{\beta}$ 的函数。在 LEF 情况下，$r(y,\mathbf{x},\boldsymbol{\theta})=(y-\mathrm{E}[y|\mathbf{x}])$（参见 5.7.3 节），因此，式(7.32)中的广义残差 $r(\cdot)$ 就是通常的残差。

更一般地讲，当 $f(y)$ 建立在两个参数密度的基础上，信息矩阵关于两个参数是分块对角的，同时，两个参数分别依赖于回归元和参数向量 $\boldsymbol{\beta}$、$\boldsymbol{\alpha}$，而 $\boldsymbol{\beta}$ 与 $\boldsymbol{\alpha}$ 却是截然不同的，此时将得到类似于式(7.32)的分解。于是，对 $\boldsymbol{\beta}$ 的 LM 检验就是对 $\tilde{\mathbf{g}}_{\beta i}$ 与 $\tilde{r}_{\beta i}$ 相关性的检验，其中，$\mathbf{s}(\boldsymbol{\beta})=\mathbf{g}_{\beta}(\mathbf{x},\boldsymbol{\theta})r_{\beta}(y,\mathbf{x},\boldsymbol{\theta})$，对 $\boldsymbol{\alpha}$ 的 LM 检验亦可给出类似解释。

一个重要例子是在正态性下具有两个参数 μ 与 σ^2 的线性回归，其中，μ 与 σ^2 被建模成 $\mu=\mathbf{x}'\boldsymbol{\beta}$ 与 $\sigma^2=\alpha$ 或 $\sigma^2=\sigma^2(\mathbf{z},\boldsymbol{\alpha})$。对正态性条件下线性回归的排除性约

束来说，$s_i(\boldsymbol{\beta})=\mathbf{x}_i(y_i-\mathbf{x}_i'\boldsymbol{\beta})$，而且 LM 检验是对回归元 $\mathbf{x}_i$ 与约束模型残差 $\tilde{u}_i=y_i-\mathbf{x}_{1i}'\tilde{\boldsymbol{\beta}}_1$ 之间相关性的检验。对具有异方差性 $\sigma_i^2=\exp(\alpha_1+\mathbf{z}_i'\boldsymbol{\alpha}_2)$ 的检验来说，$s_i(\boldsymbol{\alpha})=\frac{1}{2}\mathbf{z}_i((y_i-\mathbf{x}_i'\tilde{\boldsymbol{\beta}})^2/\tilde{\sigma}_i^2)-1)$，而拉格朗日乘子检验是对 $\mathbf{z}_i$ 与残差平方 $\tilde{u}_i^2=(y_i-\mathbf{x}_1'\tilde{\boldsymbol{\beta}})^2$ 之间相关性的检验，这是因为在零假设下，$\boldsymbol{\alpha}_2=\mathbf{0}$ 为常值。

拉格朗日乘子检验梯度形式的外积

现在回到式(7.31)所定义的一般形式 $s_i(\boldsymbol{\theta})$ 上。下面我们证明，拉格朗日乘子检验统计量(7.25)的渐近等价形式，能通过实施辅助回归(**auxiliary regression**)或人工回归：

$$1=\tilde{\mathbf{s}}_i'\boldsymbol{\gamma}+v_i \tag{7.33}$$

得到。其中，$\tilde{\mathbf{s}}_i=s_i(\tilde{\boldsymbol{\theta}}_r)$，并计算：

$$\mathrm{LM}^*=NR_u^2 \tag{7.34}$$

这里，R_u^2 表示式(7.36)后面所定义的非中心 R^2。LM^* 在 H_0 下服从 $\chi^2(h)$。等价地讲，LM^* 等于 ESS_u，即非中心解释的平方和(拟合值的平方和)；或者等于 $N-\mathrm{RSS}$，其中 RSS 源自式(7.33)中的残差平方和。

像许多应用一样，该结果很容易实施，并十分容易以解析形式得到 $s_i(\boldsymbol{\theta})$，生成 q 个分量 $\tilde{\mathbf{s}}_{1i},\cdots,\tilde{\mathbf{s}}_{qi}$ 的数据，并把 1 对 $\tilde{\mathbf{s}}_{1i},\cdots,\tilde{\mathbf{s}}_{qi}$ 进行回归。注意到，在这里，式(7.31)中的 $f(y_i|\mathbf{x}_i,\boldsymbol{\theta})$ 是无约束模型的密度。

对 7.3.2 节中泊松模型例子的排除性约束来说，$s_i(\boldsymbol{\beta})=(y_i-\exp(\mathbf{x}_i'\tilde{\boldsymbol{\beta}}))\mathbf{x}_i$ 且 $\mathbf{x}_i'\tilde{\boldsymbol{\beta}}_r=\mathbf{x}_{1i}'\tilde{\boldsymbol{\beta}}_{1r}$。由此可得，$\mathrm{LM}^*$ 被计算成从 1 对 $(y_i-\exp(\mathbf{x}_{1i}'\tilde{\boldsymbol{\beta}}_{1r}))\mathbf{x}_i$ 的回归中得到的 NR_u^2，其中，$\mathbf{x}_i$ 既包括 $\mathbf{x}_{1i}$ 又包括 $\mathbf{x}_{2i}$，而 $\tilde{\boldsymbol{\beta}}_{1r}$ 是从 y_i 对 $\mathbf{x}_{1i}$ 进行的泊松回归中得到的。

式(7.33)与式(7.34)只要求针对不同 i 的独立性。当对结果进一步假定时，可能得出其他一些辅助回归。特别地，对如同式(7.32)的 $\mathbf{s}(\boldsymbol{\theta})$ 分解因式情况专门研究，并定义 $r(y,\mathbf{x},\boldsymbol{\theta})$，故 $\mathrm{V}[r(y,\mathbf{x},\boldsymbol{\theta})]=1$。于是，拉格朗日乘子检验的一种可供选择的渐近等价形式，是来自 $\tilde{r}_i$ 对 $\tilde{\mathbf{g}}_i$ 回归的 NR_u^2。这就得出在正态情况下线性回归的拉格朗日乘子检验，譬如异方差性布鲁什—帕甘(Breusch-Pagan)LM 检验。

这些可供选择的 LM 检验形式称为拉格朗日乘子检验梯度形式的外积(**outer-product-of-the-gradient**)，因为它们通过 $\mathbf{B}_0$ 的梯度外积(OPG)估计值或 BHHH 估计值来代替式(7.22)中的 $-\mathbf{A}_0$。尽管它们很容易计算，但拉格朗日乘子检验的 OPG 变形具有不好的小样本性质，且有很大水平的扭曲。这妨碍了对拉格朗日乘子检验的 OPG 形式的运用。这些小样本问题能通过自助法(参见 11.6.3 节)而大大减少。戴维森和麦金农(Davidson and MacKinnon, 1984)提出了也可以在有限样本中较好实施的双倍长度的辅助回归。

OPG 形式的推导

为了推导 LM^*，首先，注意到式(7.25)中，$\partial\ln L(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\tilde{\theta}_r}=\sum\tilde{\mathbf{s}}_i$。其次，由信息矩阵等式 $\mathbf{A}_0=-\mathbf{B}_0$ 和 5.5.2 节可知，$\mathbf{B}_0$ 在 H_0 下通过 OPG 估计值或BHHH 估计值 $N^{-1}\sum_i\tilde{\mathbf{s}}_i\tilde{\mathbf{s}}_i'$ 得到一致估计。综合考虑这些结果，就得出拉格朗日乘子检验统

计量(7.25)的一种渐近等价形式：

$$\mathrm{LM}^{*}=\left(\sum_{i=1}^{N}\mathbf{s}_i'\right)\left[\sum_{i=1}^{N}\mathbf{s}_i\mathbf{s}_i'\right]^{-1}\left(\sum_{i=1}^{N}\mathbf{s}_i\right) \tag{7.35}$$

这一统计量能从 1 对 $\mathbf{s}_i$ 的辅助回归中计算出来，如下所示。定义 **S** 表示 $N\times q$ 阶矩阵，其中，第 i 行为 $\mathbf{s}_i'$，定义 **1** 表示元素为 1 的 $N\times 1$ 阶向量。于是，有：

$$\mathrm{LM}^{*}=\mathbf{1}'\mathbf{S}[\mathbf{S}'\mathbf{S}]^{-1}\mathbf{S}'\mathbf{1}=\mathrm{ESS}_u=NR_u^2 \tag{7.36}$$

通常，对于 **y** 对 **X** 的回归来说，非中心化解释平方和(ESS)(**uncentered explained sums of squares**)是 $\mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$，它确实是式(7.36)的形式，其中，非中心化(**uncentered**)R^2 是 $R_u^2=\mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}/\mathbf{y}'\mathbf{y}$，此处是由 $\mathbf{1}'\mathbf{1}=N$ 去除式(7.36)。之所以使用非中心化术语，是因为在 R_u^2 除法中，利用了在 0 点的而不是样本均值处的离差平方和。

7.4 例子：基于似然的假设检验

各种检验方法——沃尔德、LR 以及 LM——都利用从 $y|\mathbf{x}$ 的泊松分布数据生成过程所得到的数据加以阐述，其中，均值为 $\exp(\beta_1+\beta_2x_2+\beta_3x_3+\beta_4x_4)$，这里，$\beta_1=0$ 且 $\beta_2=\beta_3=\beta_4=0.1$，并且这三个回归元都是从 $\mathcal{N}[0,1]$iid 抽取的。

样本量为 200 的 y 对截距、x_2、x_3 以及 x_4 的泊松回归，得出无约束 MLE：

$$\hat{\mathrm{E}}[y|x]=\exp(\underset{(-2.14)}{-0.165}\underset{(-0.36)}{-0.0028x_2}\underset{(2.43)}{+0.163x_3}\underset{(0.08)}{+0.103x_4})$$

其中，有关 t 统计量已由括号给出，同时无约束对数似然是 -238.772。

四种不同的假设检验分析已详细列在表 7.1 的第 1 列里。估计量是非线性的，而其假设分别是单个排除性约束、多重排除性约束、线性约束以及非线性约束。此表的其他内容给出这四个检验的渐近等价检验统计量及其相关的 p 值。对于该样本来说，所有检验都在显著性水平 0.05 上拒绝前两个检验，而对其余两个检验则不拒绝。

表 7.1 泊松回归例子的检验统计量[a]

零假设	检验统计量				$\ln L$	水平 0.05
	沃尔德	LR	LM	LM*		的结果
$H_{10}:\beta_3=0$	5.904 (0.015)	5.754 (0.016)	5.916 (0.015)	6.218 (0.013)	−241.648	拒绝
$H_{20}:\beta_3=0,\ \beta_4=0$	8.570 (0.014)	8.302 (0.016)	8.575 (0.014)	9.186 (0.010)	−242.922	拒绝
$H_{10}:\beta_3-\beta_4=0$	0.293 (0.588)	0.293 (0.589)	0.293 (0.588)	0.315 (0.575)	−238.918	不拒绝
$H_{10}:\beta_3/\beta_4-1=0$	0.158 (0.691)	0.293 (0.589)	0.293 (0.588)	0.315 (0.575)	−238.918	不拒绝

[a] y 的数据生成过程是泊松分布，其参数为 $\exp(0.0+0.1x_2+0.1x_3+0.1x_4)$，样本量为 $N=200$。与括号中 p 值有关的检验统计量已经给出。第二个假设检验是 $\chi^2(2)$分布，而其他一些检验是 $\chi^2(1)$分布。约束 ML 估计的对数似然值也已给出；无约束模型的对数似然是 -238.772。

利用式(7.23)可计算沃尔德检验统计量。为了获得无约束 MLE 的方差矩阵估计值,需要对前面给出的无约束模型进行估计。于是,各种不同检验的沃尔德检验,需对不同的 **h** 与 **R** 进行计算,并在一些情况下加以简化。单个排除性约束的沃尔德卡方检验正是通常 t 检验的平方,即 $2.43^2 \simeq 5.90$。联合排除性约束的沃尔德检验统计量已在 7.2.5 节详细阐述。这里,x_3 是统计显著的,而 x_4 是统计不显著的,但是,x_3 与 x_4 联合在水平 0.05 上都是统计显著的。第三个假设的沃尔德检验已由式(7.19)给出,而且不能拒绝。第三个假设与第四个假设是等价的,因为 $\beta_3/\beta_4-1=0$ 蕴含 $\beta_3=\beta_4$,但第四个检验的沃尔德检验已由式(7.13)给出,不同于式(7.19)。利用矩阵运算,可计算出式(7.13)统计量,因为大部分软件包都计算线性假设的沃尔德检验。

给定约束模型的估计,利用式(7.21)特别容易计算似然比检验统计量。对前三个假设来说,约束模型是通过 y 分别对回归元$(1,x_2,x_4)$、$(1,x_2)$以及$(1,x_2,x_3+x_4)$进行泊松回归而估计出的,其中,第三个回归使用了如果 $\beta_3=\beta_4$ 则 $\beta_3x_3+\beta_4x_4=\beta_3(x_3+x_4)$的条件。举一个似然比检验的例子,对第二个假设来说,$\text{LR}=-2[-238.772-(242.922)]=8.30$。第四个约束模型原则上受限于参数为非线性约束的 ML 估计,少数几个软件包可以这样做。不过,对约束表述方式来说,受约束的 ML 估计是不变的,因此,对于第三个约束模型,可得出相同估计值,进而导致同样的 LR 检验统计量。

将泊松模型特殊化为式(7.27),利用式(7.25)计算 LM 检验统计量。该统计量利用矩阵命令来计算,各种不同的约束会得到不同的约束 MLE 估计值$\tilde{\boldsymbol{\beta}}$。如同似然比检验,LM 检验针对变换是不变的,因此,第三个假设与第四个假设的 *LM* 检验是等价的。

LM 检验统计量的渐近等价形式是式(7.35)给出的统计量 LM^*。这能够计算成为源自辅助回归(7.33)的解释平方和。对泊松模型来说,$s_{ji}=\partial\ln f(y_i)/\partial\beta_j=(y_i-\exp(\mathbf{x}_i'\boldsymbol{\beta}))x_{ji}$,对考虑的假设而言,在适当约束的 MLE 处计算。计算统计量 LM^* 比计算 LM 更容易些,尽管像 LM 一样,这需要约束 ML 估计值。

在这个带有生成数据的例子中,各种不同的检验统计量是非常相似的。情况并不总是这样的。特别地,与 LM 相比,检验统计量 LM^* 更具有不好的有限样本量性质(finite-sample size properties),即使数据生成过程是已知的。此外,在使用真实数据的应用中,数据生成过程不可能是完全设定的,甚至无穷大样本也会导致各种检验统计量的发散。

7.5 非 ML 背景下的检验

沃尔德检验是用于非 ML 背景下的一种标准检验。由 7.2 节可知,它是一种一般的检验方法,利用参数估计方差矩阵的适当三明治估计值,它总是可实施的。其唯一的局限性是,在一些应用中,实施无约束估计比实施约束估计更困难。

建立在无约束模型在约束估计值处计算出的梯度向量异于零基础上的沃尔德检验或者得分检验,同样能被推广到非 ML 估计量上。不过,通常 LM 检验的形式

比 ML 情况下的更为复杂。此外,建立在辅助回归基础上的 LM 检验的最简单形式,对错误设定分布而言不是稳健的。

似然比检验是建立在施加约束的目标函数之极大值与无约束的目标极大值之差的基础上。由于此差值通常不服从卡方分布,故除似然函数之外,这通常不能推广到目标函数上。

为了完整起见,我们提出 ML 检验推广到 m 估计量以及有效广义矩方法估计量上的概述表示。正如已注意到的,在大部分应用中,运用较简单的沃尔德检验就足够了。

7.5.1 *基于 m 估计量的检验*

对 m 估计量进行检验,就是对那些 ML 估计量的直接扩展,只是不再可能使用信息矩阵等式来简化该检验统计量,而且似然比检验在非常特殊的情况下才得以推广。所得到的检验统计量在 H_0: $\mathbf{h}(\boldsymbol{\theta})=\mathbf{0}$ 下服从渐近分布,同时在局部备择假设下服从相同的非中心卡方分布。

考察对具有一阶条件 $N^{-1}\sum_i \mathbf{s}_i(\boldsymbol{\theta})=\mathbf{0}$ 的 $Q_N(\boldsymbol{\theta})=N^{-1}\sum_i q_i(\boldsymbol{\theta})$ 求极大值的 m 估计量。定义 $q\times q$ 阶矩阵 $\mathbf{A}(\boldsymbol{\theta})=N^{-1}\sum_i \partial\mathbf{s}_i(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'$,且 $\mathbf{B}(\boldsymbol{\theta})=N^{-1}\sum_i \mathbf{s}_i(\boldsymbol{\theta})\ \mathbf{s}_i(\boldsymbol{\theta})'$,$h\times q$ 阶矩阵 $\mathbf{R}(\boldsymbol{\theta})=\partial\ln\mathbf{h}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'$。设 $\hat{\boldsymbol{\theta}}_u$ 与 $\tilde{\boldsymbol{\theta}}_r$ 分别表示无约束估计量与约束估计量,并设 $\hat{\mathbf{A}}=\mathbf{A}(\hat{\boldsymbol{\theta}}_u)$ 且 $\tilde{\mathbf{A}}=\mathbf{A}(\tilde{\boldsymbol{\theta}}_r)$,对于 $\mathbf{B}$ 与 $\mathbf{R}$ 可用类似记号。最后,设 $\hat{\mathbf{h}}=\mathbf{h}(\hat{\boldsymbol{\theta}}_u)$ 且 $\tilde{\mathbf{s}}_i=\mathbf{s}_i(\tilde{\boldsymbol{\theta}}_r)$。

沃尔德检验统计量是建立在 $\hat{\mathbf{h}}$ 接近于 0 的基础上。这里:

$$W=\hat{\mathbf{h}}[\hat{\mathbf{R}}N^{-1}\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}\hat{\mathbf{R}}']^{-1}\hat{\mathbf{h}} \tag{7.37}$$

由 5.5.1 节知,因为 $\hat{\boldsymbol{\theta}}_u$ 的稳健方差矩阵估计值是 $N^{-1}\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}$。为了计算统计推断的沃尔德检验,具有稳健标准误差项选项的软件包就运用了这种更一般形式。

设 $\mathbf{g}(\boldsymbol{\theta})=\partial\ln Q_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ 表示梯度向量,并设 $\tilde{\mathbf{g}}=\mathbf{g}(\tilde{\boldsymbol{\theta}}_r)=\sum_i\tilde{\mathbf{s}}_i$。LM 检验统计量建立在 $\tilde{\mathbf{g}}$ 接近于 0 的基础上,并且由下式给出:

$$LM=N\tilde{\mathbf{g}}'[\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{R}}'(\tilde{\mathbf{R}}\tilde{\mathbf{A}}^{-1}\hat{\mathbf{B}}\tilde{\mathbf{A}}'^{-1}\tilde{\mathbf{R}}')^{-1}\tilde{\mathbf{R}}\tilde{\mathbf{A}}^{-1}]^{-1}\tilde{\mathbf{g}} \tag{7.38}$$

通过建立基于式(7.29)的卡方检验统计量来获得此结果,其中,$N\tilde{\mathbf{g}}$ 代替 $|\partial\ln L/\partial\boldsymbol{\theta}|_{\tilde{\theta}_r}$。很明显,这个检验不像稳健的沃尔德检验那样可直接实施。8.4 节将给出稳健形式的 LM 检验的几个计算例子。在计算机软件包中,LM 检验的标准实施常常不是 LM 检验的稳健形势。

对似然比检验加以推广并不容易。若对某个纯量 α,有 $\mathbf{B}_0=-\alpha\mathbf{A}_0$,即 IM 等式的较弱形式,则将它推广到 m 估计量上。在此类特殊情况下,准似然比(QLR)检验统计量是:

$$QLR=-2N[Q_N(\tilde{\boldsymbol{\theta}}_r)-Q_N(\hat{\boldsymbol{\theta}}_u)]/\hat{\alpha}_u \tag{7.39}$$

其中,$\hat{\alpha}_u$ 表示无约束情况下获得的 α 的一致估计值[参见伍德里奇(Wooldridge, 2002,第 370 页)]。对于广义线性模型,条件 $\mathbf{B}_0=-\alpha\mathbf{A}_0$ 成立(参见 5.7.4 节)。于是,统计量准似然比等价于约束模型与无约束模型偏差之差,即建立在对于 OLS 和具有同方差误差的 NLS 估计来说的约束残差平方和与无约束残差平方和之差

基础上的一般化 F 检验。对一般的准 ML 估计来说，满足 $\mathbf{B}_0 \neq -\alpha\mathbf{A}_0$，似然比检验统计量服从加权卡方和分布(参见 8.5.3 节)。

7.5.2 建立在有效 GMM 估计量基础上的检验

对于广义矩方法，各种检验统计量就有效广义矩方法而言是最简单的，这意味着广义矩方法估计利用了最优加权矩阵。由于总可以估计出最优加权矩阵，其详细内容如 6.3.5 节所述，因而这并没有对实际应用产生很大束缚。

考察建立在矩条件 $\mathrm{E}[\mathbf{m}_i(\boldsymbol{\theta})]=\mathbf{0}$ 基础上的广义矩方法估计。[注意，第 6 章的记号在这里有些变化：本章用 $\mathbf{h}(\boldsymbol{\theta})$ 表示在 H_0 下的约束。]当利用 6.3.5 节引入记号时，有效的无约束广义矩方法估计量 $\hat{\boldsymbol{\theta}}_u$ 对 $Q_N(\boldsymbol{\theta})=\mathbf{g}_N(\boldsymbol{\theta})'\ \mathbf{S}_N^{-1}\mathbf{g}_N(\boldsymbol{\theta})$ 求极小值，其中，$\mathbf{g}_N(\boldsymbol{\theta})=N^{-1}\sum_i \mathbf{m}_i(\boldsymbol{\theta})$，并且 $\mathbf{S}_N$ 关于 $\mathbf{S}_0=\mathrm{V}[\mathbf{g}_N(\boldsymbol{\theta})]$ 是一致的。约束广义矩方法的估计量 $\tilde{\boldsymbol{\theta}}_r$ 被假定成，对具有相同加权矩阵 $\mathbf{S}_N^{-1}$ 的 $Q_N(\boldsymbol{\theta})$ 求极小值，使得约束 $\mathbf{h}(\boldsymbol{\theta})=\mathbf{0}$。

纽韦和韦斯特(Newey and West, 1987a)曾经总结了下述三个检验统计量在 H_0：$\mathbf{h}(\boldsymbol{\theta})=\mathbf{0}$ 下都渐近服从 $\chi^2(h)$ 分布，且在局部备择假设下服从相同的非中心卡方分布。

与以往一样，沃尔德检验统计量建立在 $\hat{\mathbf{h}}$ 接近于 0 的基础上。这就得到：

$$\mathrm{W}=\hat{\mathbf{h}}'[\hat{\mathbf{R}}N^{-1}(\hat{\mathbf{G}}'\mathbf{S}^{-1}\hat{\mathbf{G}})^{-1}\hat{\mathbf{R}}']\hat{\mathbf{h}} \tag{7.40}$$

因为由 6.3.5 节知，有效广义矩方法估计量的方差为 $N^{-1}(\hat{\mathbf{G}}'\mathbf{S}^{-1}\hat{\mathbf{G}})^{-1}$，其中，$\mathbf{G}_N(\boldsymbol{\theta})=\partial\mathbf{g}_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'$，并且 ^表示在 $\hat{\boldsymbol{\theta}}_u$ 处所计算的值。

有效广义矩方法的一阶条件是 $\hat{\mathbf{G}}'\mathbf{S}^{-1}\hat{\mathbf{g}}=\mathbf{0}$。不过。当在 $\tilde{\boldsymbol{\theta}}_r$ 处计算时，LM 统计量检验了这个梯度向量是否接近于 0，得出：

$$\mathrm{LM}=N\,\tilde{\mathbf{g}}'\mathbf{S}^{-1}\tilde{\mathbf{G}}(\tilde{\mathbf{G}}'\mathbf{S}^{-1}\tilde{\mathbf{G}})^{-1}\tilde{\mathbf{G}}'\mathbf{S}^{-1}\tilde{\mathbf{g}} \tag{7.41}$$

其中，~表示在 $\tilde{\boldsymbol{\theta}}_r$ 处计算的值，同时利用 6.3.3 节的假设：$\sqrt{N}\mathbf{g}_N(\boldsymbol{\theta}_0)\xrightarrow{d}\mathcal{N}[0,\mathbf{S}_0]$，故 $\sqrt{N\mathbf{G}'\mathbf{S}^{-1}\mathbf{g}}\xrightarrow{d}\mathcal{N}[\mathbf{0},\operatorname{plim}N^{-1}\mathbf{G}'\mathbf{S}^{-1}\mathbf{G}]$。

对有效广义矩方法估计量来说，对目标函数的极大值方面的差异进行比较，从而得出差分检验统计量(difference test statistic)：

$$\mathrm{D}=N[Q_N(\tilde{\boldsymbol{\theta}}_r)-Q_N(\hat{\boldsymbol{\theta}}_u)] \tag{7.42}$$

如同 W 与 LM，统计量 D 在 H_0 下渐近服从 $\chi^2(h)$ 分布。

甚至在似然情况下，这个最后的统计量不同于似然比统计量，因为它使用了不同的目标函数。MLE 使 $Q_N(\boldsymbol{\theta})=-N^{-1}\sum_i \ln f(y_i|\boldsymbol{\theta})$ 极小化。由 6.3.7 节可知，相反，渐近等价的有效广义矩方法估计量对二次形式 $Q_N(\boldsymbol{\theta})=N^{-1}(\sum_i s_i(\boldsymbol{\theta}))'\times(\sum_i s_i(\boldsymbol{\theta}))$ 求极小值，其中，$s_i(\boldsymbol{\theta})=\partial\ln f(y_i|\boldsymbol{\theta})/\partial\boldsymbol{\theta}$。倘若所用的广义矩方法估计量是有效广义矩方法估计量，通常就可以运用统计量 D，而似然比检验则只有在式(7.39)后面所提及的 m 估计量的某些特殊情况下才得以推广。

对于矩方法估计量，也就是说，在恰好识别广义矩方法模型中，D＝LM＝

$NQ_N(\tilde{\boldsymbol{\theta}}_r)$，因此，LM与差检验是等价的。就D而言，出现简化，因为$g_N(\hat{\boldsymbol{\theta}}_u)=\mathbf{0}$，从而$Q_N(\hat{\boldsymbol{\theta}}_u)=0$。对于LM，会使式(7.41)简化，就如同$\tilde{\mathbf{G}}_N$是可逆的一样。

7.6 检验势与水平

本章余下几节研究，运用通常的计算机输出进行假设检验的两个局限性。

首先，一个检验很少有能力去区别零假设与备择假设。于是，该检验具有低的势，意味着当零假设是错误的时候，拒绝零假设的概率很小。标准的计算机输出不会计算检验的势，但计算机能利用渐近方法(参见本节)或有限样本蒙特卡罗方法(参见7.7节)加以计算。当经验论文的主要贡献是对特殊假设进行拒绝或者没有拒绝，该论文就没有理由去另外阐述针对某个有意义备择假设检验的势。

其次，检验的真实水平本质上可以不同于由渐近理论得出的检验名义水平。一种经验做法是：为了得到单变量推断的良好近似，对渐近理论来说，样本量$N>30$就足够了，但经验做法不能推广到具有回归元的模型上。不好的近似可能出现在逼近分布的尾部，但其尾部经常用于获得通常诸如5%显著性水平上检验的临界值。实际上，对从大样本近似中获得的检验统计量来说，其临界值常常小于建立在未知真实分布基础上的正确临界值。小样本精炼企图得到更接近于准确的临界值。对线性回归来说，在正态情况下，得到准确的临界值，利用t而不是z，以及F而不是χ^2分布，可是就非线性回归而言，其类似结果将准确。不过，通过蒙特卡罗方法(参见7.7节)或利用自助法(参见7.8节与第11章)，可使小样本精炼。

借助于现代计算机，对应用研究所用到的检验水平进行修正并研究其检验的势，相对很容易。我们以某种详细方式来阐述这个被忽略的专题。

7.6.1 *检验水平与势*

假设检验导致对零假设的拒绝，或者导致对零假设的不拒绝。当H_0不正确时拒绝H_0，或者当H_0正确时没有拒绝H_0，这些都是正确决策。

同样存在两种可能的错误决策：(1)当H_0是正确时，拒绝H_0，称之为第Ⅰ类错误(**type I error**)；(2)当H_0是错误时，没有拒绝H_0，称之为第Ⅱ类错误(**type II error**)。理想地说，这两类错误的概率都很小；但实际上，一类错误的概率减少会以另一类错误的概率增大为代价。经典假设检验求解方法就是将第Ⅰ类错误概率固定在某个特殊水平上，通常是0.05，而对第Ⅱ类错误不进行设定。

定义检验水平(**size of a test**)或显著性水平(**significance level**)为：

$$\begin{aligned}\alpha &= \Pr[\text{第Ⅰ类错误}] \\ &= \Pr[\text{拒绝 } H_0 \mid H_0 \text{ 正确}]\end{aligned} \tag{7.43}$$

对α普遍选取为0.01、0.05或0.10。当该检验统计量落入所定义的拒绝域中，就拒绝假设，因而检验的显著性水平等于设定的α值。与之密切相关的一种等价方法是，计算检验p值(***p*-value**)，在p值的临界显著性水平上，零假设刚好被拒绝；而当p值小于所设定α值时，就拒绝H_0。这两种方法只需要知道检验统计量在零假

设下的分布知识，7.2 节已经对沃尔德检验统计量进行了阐述。

还应该考虑给出第Ⅱ类错误的概率。定义检验势(**power of a test**)为：

$$\begin{aligned}\text{势} &= \Pr[\text{拒绝 } H_0 \mid H_0 \text{ 正确}] \\ &= 1-\Pr[\text{接受 } H_0 \mid H_0 \text{正确}] \\ &= 1-\Pr[\text{第 Ⅱ 类错误}]\end{aligned} \tag{7.44}$$

在理论上，检验势接近于 1，这是因为第Ⅱ类错误概率接近于 0。想要确定势，就需要检验统计量在 H_a 下的分布知识。

在实证研究中，对检验的势所进行的分析典型地被忽略了，给定水平 α，检验方法通常被选择为在理论上具有势的情况除外，势与其他备择的检验统计量高度相关联。从理论上看，可使用一致最大势(**uniformly most powerful, UMP**)检验。当对简单零假设对应于简单备择假设进行检验时，一致最大势检验就会存在。于是，内曼一皮尔逊引理给出一致最大势检验是似然比的函数这个结论。对于涉及复合假设的更一般检验情况来说，通常不存在一致最大势检验，而且可设置诸如一致最大势单边检验的进一步约束。实际上，把对势的考察留给理论经济计量学家，他们使用理论及模拟应用到各种检验方法上，以此确立哪一种检验方法的势最大。

不过，在任何已知应用中，有可能决定检验的势。下面详述如何计算沃尔德检验的渐近势，它等于完全参数情况下的 LR 检验与 LM 检验的渐近势。

7.6.2 *局部备择假设*

当 H_a 是正确时，由于势是拒绝 H_0 的概率，所以对势进行计算需要获得检验统计量在备择假设下的分布。对显著性水平 α 的沃尔德卡方检验来说，其势等于 $\Pr[W>\chi^2_\alpha(h) \mid H_a]$。计算这个概率需要对特定的备择假设加以设定，因为 H_a：$\mathbf{h}(\boldsymbol{\theta})\neq\mathbf{0}$ 是非常广泛的。

一种明显的选择是固定备择假设 $\mathbf{h}(\boldsymbol{\theta})=\boldsymbol{\delta}$，其中，$\boldsymbol{\delta}$ 表示有限非零常值的向量。有时候，数量 $\boldsymbol{\delta}$ 称为假设误差，而且较大的假设误差会导致较大的势。对固定备择假设来说，沃尔德检验统计量渐近具有势 1，因为它始终拒绝零假设。为了理解这一点，注意到，当 $\mathbf{h}(\boldsymbol{\theta})=\boldsymbol{\delta}$，沃尔德检验统计量变成无限的，因为：

$$\begin{aligned}W &= \hat{\mathbf{h}}'(\hat{\mathbf{R}}N^{-1}\hat{\mathbf{C}}\hat{\mathbf{R}}')^{-1}\hat{\mathbf{h}} \\ &\xrightarrow{p} \boldsymbol{\delta}'(\mathbf{R}_0N^{-1}\mathbf{C}_0\mathbf{R}_0')^{-1}\boldsymbol{\delta}\end{aligned}$$

利用 $\hat{\boldsymbol{\theta}}\xrightarrow{p}\boldsymbol{\theta}_0$，因此，$\hat{\mathbf{h}}=\mathbf{h}(\hat{\boldsymbol{\theta}}_u)\xrightarrow{p}\mathbf{h}(\boldsymbol{\theta})=\boldsymbol{\delta}$，且 $\hat{\mathbf{C}}\xrightarrow{p}\mathbf{C}_0$。由此可得，因为除 N 之外所有项都是有限的且非零的，故 $W\xrightarrow{p}\infty$。该无限值会使 H_0 总是被拒绝，因为它的势为 1，进而具有完全势 1。

因此，沃尔德检验统计量是一致检验统计量(**consistent test statistic**)，也就是说，其势随着 $N\to\infty$ 而趋于 1。许多检验统计量是一致的，正如许多估计量是一致的一样。为了区分一些检验统计量，需要更为严格的准则，正如相对有效性用于选择估计量一样。

对于作为根－N（root-N）一致的估计量来说，我们考虑局部备择序列（**sequence of local alternatives**）：

$$H_a: \mathbf{h}(\boldsymbol{\theta})=\boldsymbol{\delta}/\sqrt{N} \quad (7.45)$$

其中，$\boldsymbol{\delta}$ 表示固定常值的向量，满足 $\boldsymbol{\delta}\neq\mathbf{0}$。这种备择假设序列被称为皮特曼漂移（**Pitman drift**），它因为样本量愈大而愈接近于零假设的零值，以相同的速率$\sqrt{N}$用作对 $\hat{\boldsymbol{\theta}}$ 的标度，获得一致估计量的非退化分布。因此，$\mathbf{h}(\boldsymbol{\theta})$的备择假设值以一种使随着样本增加，任何改进有效性都无效的速率趋于零。有关对局部备择给出的更详细的解释及文献，可参见麦克马纳斯（MacManus，1991）。

7.6.3 *沃尔德检验渐近势*

在局部备择结果（7.45）下，沃尔德检验统计量具有非退化分布、非中心卡方分布。这使得确定沃尔德检验的势成为可能。

特别地，正如7.6.4节将证明的，在 H_a 下，式（7.6）所定义的沃尔德统计量渐近服从 $\chi^2(h;\lambda)$分布，其中，$\chi^2(h;\lambda)$表示非中心卡方分布（**noncentral chi-square distribution**），该非中心参数（**noncentrality parameter**）为：

$$\lambda=\frac{1}{2}\boldsymbol{\delta}'(\mathbf{R}_0\mathbf{C}_0\mathbf{R}_0')^{-1}\boldsymbol{\delta} \quad (7.46)$$

其中，$\mathbf{R}_0$ 与 $\mathbf{C}_0$ 已由式（7.4）与式（7.5）定义。因此，沃尔德检验的势（**Wald test of power**）是给定局部备择 H_a 为正确时拒绝 H_0 的概率，即：

$$\text{势}=\Pr[W>\chi_\alpha^2(h)\,|\,W\sim\chi_\alpha^2(h;\lambda)] \quad (7.47)$$

图7.1画出当水平或显著性水平分别为10%、5%、1%时，广泛运用的对纯量假设（$h=1$）进行检验的 λ 势。当 λ 接近于1时，其势等于水平；而对于大 λ 而言，其势接近于1。这些特性对 $h>1$ 也成立。特别是，势关于式（7.46）所定义的非中心性参数 λ 是单调递增的。后面将阐述几个一般性结果。

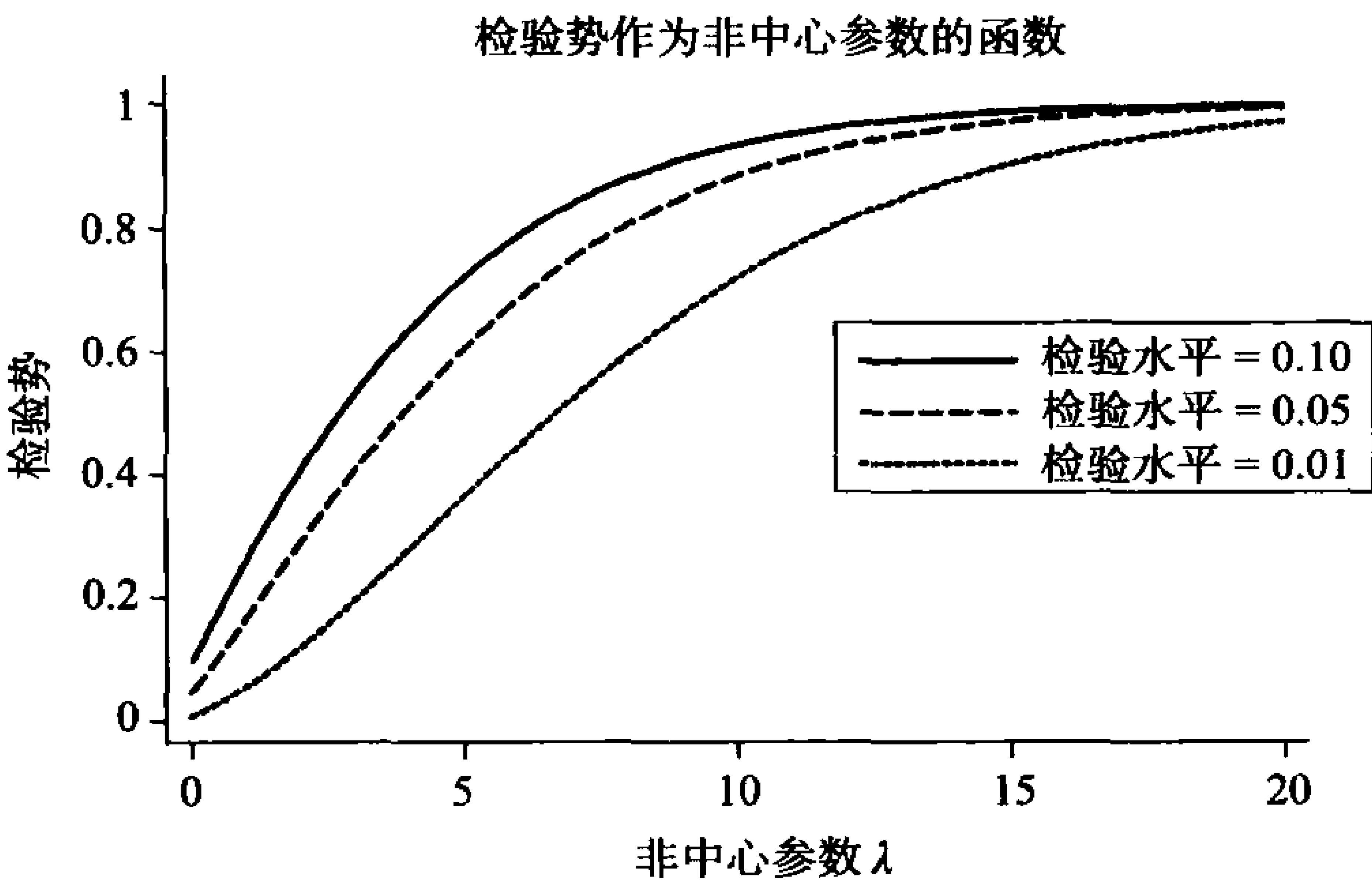


图7.1 当非中心参数从0到20变动时，具有一个自由度、三种不同检验水平的沃尔德卡方检验的势。

第一，势关于零假设与备择假设之间的距离是递增的，进而 $\boldsymbol{\delta}$ 及 λ 都是递增的。

第二，对于给定备择 $\boldsymbol{\delta}$，势随着估计量 $\hat{\boldsymbol{\theta}}$ 的有效性而增大，进而 $\mathbf{C}_0$ 变得较小，因此，λ 会比较大。

第三，当检验水平增大时，势会变大，而第二类错误的概率则会减小。

第四，如果零假设下的几种不同检验统计量都服从 $\chi^2(h)$，且在备择假设下都服从非中心 $\chi^2(h)$，那么较受欢迎的检验统计量是带有最高(大)非中心参数 λ 的那一种，从而势是最大的。进一步地，具有相同非中心参数的两个检验在局部备择假设下是渐近等价的。

最后，在实际应用中，人们把势计算成 δ 的函数。特别地，对于设定的备择 $\boldsymbol{\delta}$，利用与 $\hat{\mathbf{R}}$ 及 $\hat{\mathbf{C}}$ 有关的参数估计值 $\hat{\boldsymbol{\theta}}$ 。使用式(7.46)就能计算估计非中心参数 $\hat{\lambda}$。7.6.5 节将阐述这种势的计算。

7.6.4 *渐近势的推导*

为了获得 H_a 下的 W 分布，以泰勒级数展开式结果(7.9)开始。在 H_a 下，这被简化为：

$$\sqrt{N}\mathbf{h}(\hat{\boldsymbol{\theta}}) \xrightarrow{d} \mathcal{N}[\boldsymbol{\delta}, \mathbf{R}_0\mathbf{C}_0\mathbf{R}_0'] \tag{7.48}$$

从而$\sqrt{N}\mathbf{h}(\boldsymbol{\theta})=\boldsymbol{\delta}$。因而，以 $\boldsymbol{\delta}$ 为中心的二次形式在 H_a 下服从卡方分布。

相反，由式(7.6)定义的沃尔德检验统计量，形成了以 0 为中心的二次形式，并在 H_a 下不再服从卡方分布。通常，若 $\mathbf{z}\sim\mathcal{N}[\boldsymbol{\mu},\boldsymbol{\Omega}]$，其中，$\text{rank}(\boldsymbol{\Omega})=h$，则 $\mathbf{z}'\boldsymbol{\Omega}^{-1}\mathbf{z}\sim\chi^2(h;\lambda)$，其中，$\chi^2(h;\lambda)$表示具有非中心性参数 $\lambda=\frac{1}{2}\boldsymbol{\mu}'\boldsymbol{\Omega}^{-1}\boldsymbol{\mu}$ 的非中心卡方分布。将这一结果应用到式(7.48)，则在 H_a 下，得出：

$$N\mathbf{h}(\hat{\boldsymbol{\theta}})'(\mathbf{R}_0\mathbf{C}_0\mathbf{R}_0')^{-1}\mathbf{h}(\hat{\boldsymbol{\theta}}) \xrightarrow{d} \chi^2(h;\lambda) \tag{7.49}$$

其中，λ 已由式(7.49)定义。

7.6.5 *渐近势的计算*

为了阐明势是如何随 $\boldsymbol{\delta}$ 而变化的，考察纯量情况下对系数显著性的检验。于是，式(7.46)定义的非中心性参数是：

$$\lambda=\frac{\delta^2}{2c}\simeq\frac{(\delta/\sqrt{N})^2}{2(\text{se}[\hat{\theta}])^2} \tag{7.50}$$

这种近似产生于 $N(\text{se}[\hat{\theta}])^2$ 对 c 的估计，即$\sqrt{N}(\hat{\theta}-\theta)$的方差极限，这里，$\text{se}[\hat{\theta}]$表示 $\hat{\theta}$ 的标准误差。

考察 H_0：$\theta=0$ 的沃尔德卡方检验，其备择检验为：θ 位于 0 的一个标准误差之内，即：

$$H_a: \theta=a\times\text{se}[\hat{\theta}]$$

这里，把 se$[\hat{\theta}]$处理成一个常值。从而，式(7.45)的 $\delta/\sqrt{N}$等于 $a\times$se$[\hat{\theta}]$，由此式(7.50)简化成 $\lambda=a^2/2$。因此，沃尔德检验在 H_a 下渐近服从 $\chi^2(1;\lambda)$，其中，$\lambda=a^2/2$。

由图 7.1 知，很明显，对于普遍的 5%显著性水平检验的情况来说，当 $a=2$ 时，其势小于 0.5；当 $a=4$ 时，其势在 0.5 左右；而当 $a=6$ 时，其势仍小于 0.9。因此，对备择假设表示成为源于 0 的许多标准误差而言，不明确的统计显著性检验能够具有低的势。从直观上讲，若 $\hat{\theta}=2$se$[\hat{\theta}]$，则 $\theta=0$ 对 $\theta=4$se$[\hat{\theta}]$的检验大致具有 0.5 的势，因为 θ 的 95%置信区间大约是(0,4se$[\hat{\theta}]$)，这意味着，$\theta=0$ 或 $\theta=4$se$[\hat{\theta}]$ 的值是可能的。

举一个更具体的例子，假定对 θ 测量由于培训项目而使工资上涨的百分率，研究发现，$\hat{\theta}=6$，其中 se$[\hat{\theta}]=4$。于是，在显著性水平 5%上，沃尔德检验没有拒绝 H_0，这是因为 $W=(6/4)^2=2.25<\chi^2_{0.05}(1)=3.96$。该项研究结论表明，培训项目并不是统计显著的。不过，人们不应该把这一点解释为如下含义：当这种检验具有低的势时，培训项目没有什么效果，这种情况具有很高的概率。例如，前面分析表明，$H_0:\theta=0$ 检验对 $H_a:\theta=16$，即相对大的培训效果具有仅为 0.5 的势，因为 $4\times$se$[\hat{\theta}]=16$。产生低势的原因包括：小样本量、大的模型误差方差以及回归元变动幅度小。

在简单情况下，为了达到一个给定的想要的势水平，可能需要求解估计最小样本量的逆问题。这种方法在医学研究中尤其流行。

安德鲁斯(Andrews，1989)为了确定在实证背景下的参数空间区域，对哪一个检验可能具有低的势，给出了利用非中心性参数的更正式的研究。他提供了许多应用例子，这些例子很容易确定，对于有意义的备择假设，这些检验具有低势。

7.7 蒙特卡罗研究

迄今为止，我们讨论的统计推断均依赖于渐近结果。对小样本来说，除了在正态性条件下对线性回归模型的线性约束进行检验之外，可利用的解析结果几乎很少。尽管如此，小样本结果却能通过蒙特卡罗研究来获得。

7.7.1 *概述*

下面是一个检验统计量的小样本性质的蒙特卡罗研究(**Monte Carlo study**)的例子。比如说，设样本量 N 为 40，并在 H_0 模型下随机生成容量为 40 的 10 000 个样本。对于每一个复制(样本)都可构成关注的检验统计量以及检验 H_0，当检验统计量落在拒绝区域中，就拒绝 H_0，这通常利用渐近结果来加以确定。

检验统计量的真实水平(**true size**)或实际水平(**actual size**)，正是复制中那些落入拒绝区域的检验统计量的部分。从理论上讲，这接近于名义水平(**nominal size**)，即对检验选取的显著性水平。例如，若检验在 5%名义检验水平上是 0.05，则希望真实水平接近于 0.05。

要确定小样本中的检验势，就需要额外模拟，其样本是该模型位于复合备择假

设 H_a 在可能模型的一个或多个特殊设定下生成的。将势计算成为复制中拒绝零假设的部分，或者利用相同的检验作为确定真实水平，或者利用拒绝域的检验的校正水平形式(**size-corrected version**)，使名义水平等于真实水平。

对蒙特卡罗研究可直接实施，但设计好的蒙特卡罗研究却存在着许多微妙差别。一个极好的讨论，参见戴维森和麦金农(Davidson and MacKinnon, 1993)。

7.7.2 蒙特卡罗内容

举一个蒙特卡罗研究的例子，我们考虑 probit 模型中对斜率系数的统计推断。下面的分析并不依赖于 probit 模型的知识。

数据生成过程是 probit 模型，二值回归元 y 以概率

$$\Pr[y=1|\mathbf{x}]=\Phi(\beta_1+\beta_2 x)$$

等于 1，其中，$\Phi(\cdot)$表示标准正态 cdf，$x\sim\mathcal{N}[0,1]$，并且$(\beta_1,\beta_2)=(1, 2)$。

对于该数据生成过程，很容易生成数据 (y,x)。首先，回归元 x 是从标准正态分布中随机抽取的。于是，由 14.4.2 节知，当 $x+u>0$，对因变量 y 设置为 1，否则设置为 0，其中，u 表示从标准正态分布中随机抽取。对该数据生成过程来说，有一半时间 $y=1$，而另一半时间 $y=0$。

在每一次模拟中，都要抽取 x 与 y 的 N 个新观测值，并从 y 对 x 的 probit 回归中获得 MLE。一种可选择的方式是，在每次模仿中，都使用相同的回归元 x 的 N 个抽取，然后再抽取 y。前一个方案对应于简单随机抽样，而后一个方案则对应于以 x 为条件的分析，或"重复试验中固定的"分析，参见 4.4.7 节。

蒙特卡罗研究经常考察一系列的样本量。这里，我们简单地设 $N=40$。也可以通过设非常大的 N 来检验项目，比如说 $N=10\,000$，从而蒙特卡罗结果应非常接近于渐近结果。

为了确定实际检验水平，需要进行大量模拟，因为这要依赖于分布的尾部而不是中心的特性。为了对真实水平 α 进行检验，而执行 S 次模拟，那么零假设被正确拒绝的次数比例是源自 S 次二项试验的结果，其二项式的均值为 α，而方差为 $\alpha(1-\alpha)/S$。因此，95%的蒙特卡罗会估计出检验水平于 $\alpha\pm1.96\sqrt{\alpha(1-\alpha)/S}$ 之间。由于仅 100 次模拟是不够的，例如，当 $\alpha=0.05$ 时，这一区间为(0.007, 0.093)。对 10 000 次模拟而言，95%的区间是更准确的，当 α 分别等于 0.01、0.05、0.10、0.20 时，该区间分别等于(0.008, 0.012)、(0.046, 0.054)、(0.094, 0.106)以及(0.192, 0.208)。这里使用 $S=10\,000$ 次模拟。

运用蒙特卡罗模拟研究时所产生的一个问题是，对某个模拟样本而言，模型可能是不可估计的。例如，考察只有一个截距与一个标示变量的线性回归。如果标示变量在模拟样本中恰好总取同一个值，比如说 0，那么它的系数就不能单独地从截距那里加以识别。在模拟样本中，当所有的 y 值都为 0 或所有的 y 值都为 1，probit 模型与其他二值结果模型就会产生类似问题。一种标准方法是要省略这种模拟样本，同时写出当出现这种问题时允许模拟循环计算的计算机编码，但人们对该方法持批评观点。在该例中，就 $N=40$ 而言，不会产生此类问题，然而，当$N=30$

时则会出现此类问题。

7.7.3 小样本偏倚

在离开检验内容之前，考察 MLE $\hat{\beta}_2$ 的小样本性质及其估计的标准误差 $\mathrm{se}[\hat{\beta}_2]$。

通过 10 000 次模拟，$\hat{\beta}_2$ 具有均值 1.201 且标准差 0.452，而 $\mathrm{se}[\hat{\beta}_2]$ 具有均值 0.359。因此，小样本的 MLE 有向上的偏倚，这是因为 $\hat{\beta}_2$ 平均远远大于 $\beta_2=1$。由于 $\mathrm{se}[\hat{\beta}_2]$ 平均远远小于 $\hat{\beta}_2$ 标准差，故小样本的标准误差是向下偏倚的。

7.7.4 检验水平

考察 $H_0: \beta_2=1$ 对 $H_a: \beta_2\neq 1$ 的双侧检验，利用沃尔德检验：

$$z=\mathrm{W}_z=\frac{\hat{\beta}_2-1}{\mathrm{se}[\hat{\beta}_2]}$$

其中，$\mathrm{se}[\hat{\beta}_2]$ 表示利用 14.3.2 节给出的方差矩阵估计出的 MLE 标准误差，它是负的期望海赛矩阵的逆。给定 dgp，从渐近形式上看，z 服从标准正态分布，且 z^2 服从卡方分布。另外，目标是求出这种如何更好地逼近小样本的分布。

图 7.2 给出 $S=10\,000$ 时计算 z 值的密度，其中的密度是利用第 9 章的核密度估计值，而不是从直方图上画出的。这增大了标准正态密度。很明显，渐近结果是不准确的，尤其在上尾部分，比如说当在 5%水平上进行检验时，其差异显然大到足以导致水平扭曲。同理，通过模拟，z 具有均值 $0.114\neq 0$ 且标准差 $0.956\neq 1$。

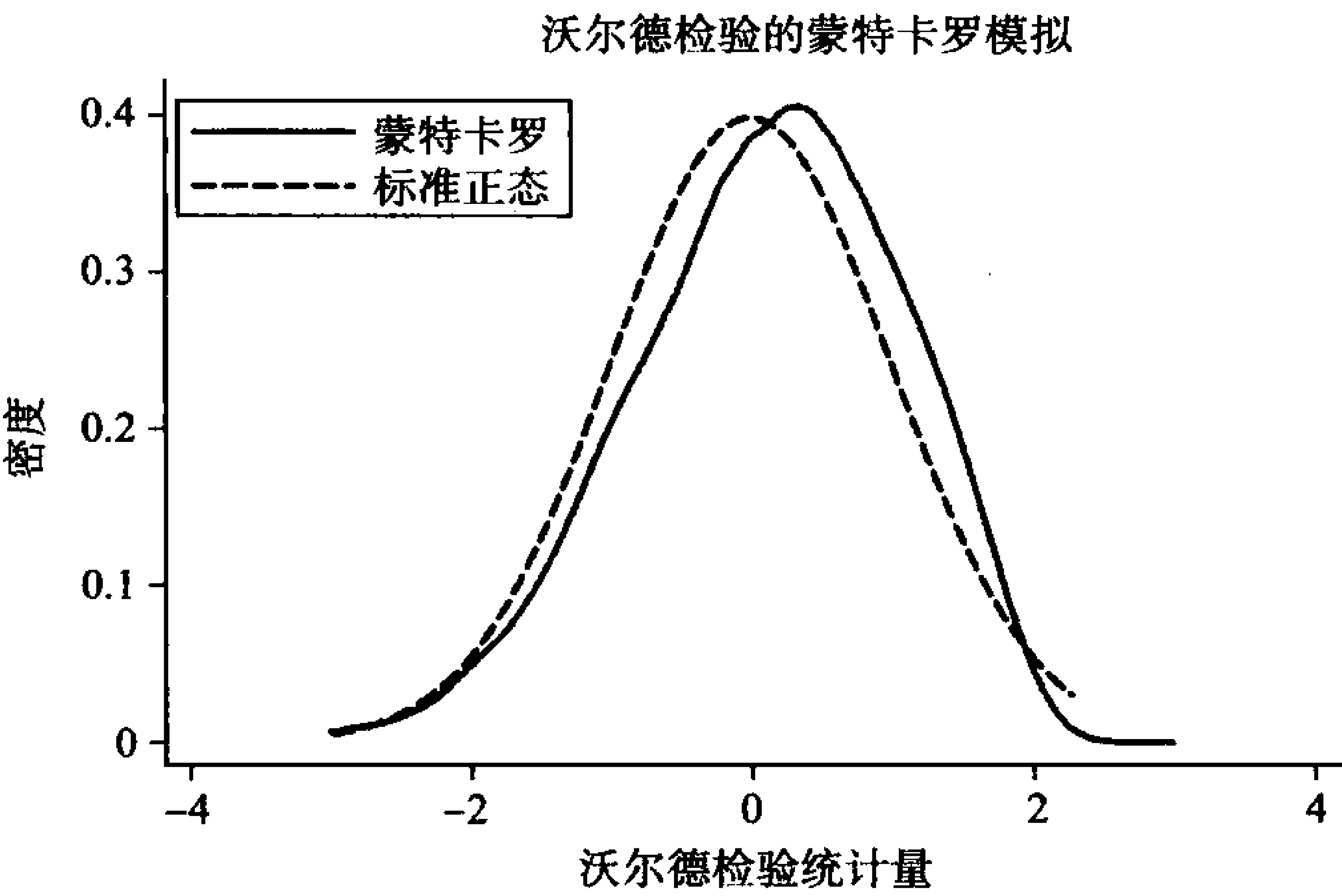


图 7.2 为了对比，此图还画出沃尔德检验统计量的密度：斜率系数等于通过带有标准正态密度的蒙特卡罗模拟而计算出的值。数据是由 probit 回归模型生成的。

表 7.2 的前两列给出沃尔德检验名义水平 α 为 0.01、0.05、0.10 和 0.20 时的名义水平与实际水平。实际水平是 10 000 次模拟满足 $|z|>z_{\alpha/2}$ 的比例，或等价地满足 $z^2>\chi_\alpha^2(1)$ 的比例。很明显，当 $\alpha\leqslant 0.1$ 时，检验的实际水平远远小于名义水平；相反，特别小样本校正假定 z 服从自由度为 38 的 t 分布，且当 $|z|>t_{\alpha/2}(38)$ 时

就要加以拒绝。不过,得出了甚至更小的实际水平,这是因为 $t_{\alpha/2}(38)>z_{\alpha/2}$。

表 7.2 probit 回归的沃尔德检验水平与势的例子[a]

名义水平 (α)	实际水平	实际势	渐近势
0.01	0.005	0.007	0.272
0.05	0.029	0.226	0.504
0.10	0.081	0.608	0.628
0.20	0.192	0.858	0.755

[a] y 的数据生成过程是满足 $\Pr[y=1|]=\Phi(0+\beta_2 x)$ 且样本量 $N=40$ 的 probit。该检验是关于斜率系数是否等于 1 的双侧沃尔德检验。实际水平是源于 $S=10\ 000$ 次模拟并满足 $\beta_2=1$ 计算出来的,而势是源于 10 000 次模拟并满足 $\beta_2=2$ 计算出来的。

蒙特卡罗模拟还能用于获得校正水平的临界值。因而,10 000 次 z 的模拟值的下 2.5%分位数与上 2.5%分位数分别是-1.905 与-2.003。由此可知,满足实际水平 0.05 的非对称拒绝域是 $z<-1.905$ 与 $z>2.003$,与 $|z_2|>1.960$ 相比,该拒绝域较大。

7.7.5 *检验势*

考察 $H_0: \beta_2=2$ 条件下的沃尔德检验势。给定 $se[\hat{\beta}_2]$具有平均值 0.359,我们希望势是合理的,因为 β_2 这个值位于远离零假设 $\beta_2=1$ 的 2～3 个标准误差。表 7.2 的最后两列给出沃尔德检验的实际势与名义势。

实际势可通过与实际水平相同的方式来获得,即 10 000 次模拟中满足 $|z|>z_{\alpha/2}$的比例。其唯一变化是,生成 y 的模拟中,$\beta_2=2$ 而不是 1。当 $\alpha=0.01$ 或 0.05,即实际水平尤其小于名义水平时,实际势是非常低的。

沃尔德检验的名义水平是利用 H_a 条件下渐近非中心 $\chi^2(1,\lambda)$分布来确定的,由式(7.50)知,$\lambda=\frac{1}{2}(\delta/\sqrt{N})^2/se[\hat{\beta}_2]^2=\frac{1}{2}\times 1^2/0.359^2\simeq 3.88$,由于局部备择假设是 $H_a: \beta_2-1=\delta/\sqrt{N}$,因而对 $\beta_2=2$ 而言,$\delta/\sqrt{N}=1$,该渐近结果并不精确,它却提供关于 $\alpha=0.10$ 与 0.20 的一个有用势估计值,即真实水平紧密地与名义水平相匹配。

7.7.6 *蒙特卡罗应用*

上面的讨论强调了运用蒙特卡罗分析时要计算检验的势与水平。通过令 N 很大,蒙特卡罗分析对求出估计量的小样本偏倚以及确定估计量是一致的,实际上也是相当有用的。利用当今的计算机软件,这类蒙特卡罗方案非常容易实施。

当给定 $\mathbf{x}$ 时 y 的条件分布是完全参数化时,可对实际数据进行蒙特卡罗分析。例如,考察具有实际数据的 probit 模型估计。在每一次模拟时,回归元都要在样本值处加以设置,如果抽样框是重复样本中固定回归元的情形,就需要生成二值因变量 y 的新值。这将依赖于使用的参数 $\boldsymbol{\beta}$ 值。设 $\hat{\beta}_1,\cdots,\hat{\beta}_k$ 表示来自原先样本的 probit 估计值,并考察 $H_0: \beta_j=0$ 的沃尔德检验。为了计算检验水平,对于$j\neq k$,通过令 $\beta_k=\hat{\beta}_k$ 且 $\beta_j=0$ 来生成 S 模拟样本,然后计算模拟的 $H_0: \beta_j=0$ 被拒绝的比

例。为了估计对特定备择假设 H_a：$\beta_j=1$ 的沃尔德检验的势，比如说，在生成 y 时，生成满足对于 $j\neq k$ 有 $\beta_k=\hat{\beta}_k$ 且 $\beta_j=1$ 的 y，同时计算模拟的 H_0：$\beta_j=0$ 被拒绝的比例。

在实际应用中，相当多的微观经济计量分析是基于估计量的，而不是建立在完全参数模拟的基础上。于是，为了实施蒙特卡罗分析，需要额外分布假设。

或者，势可用渐近方法而不是有限样本方法来获得。此外，下一节阐述的自助法能用于，通过更精致的渐近理论来得到水平。

7.8 自助法例子

自助法是蒙特卡罗模拟的一种变形，因为此种模拟具有较少的参数假设和较少的额外编程等引人注目之处，这超出了估计模型首先要求的程度。为使自助法的基本成分有效，估计量要确实服从极限分布，同时自助法再抽样的量是 iid 的。

自助法具有两种一般性应用。第一种应用是，自助法能用作一种可供选择的方式，来计算没有渐近精炼时的统计量。当解析公式很复杂时，这尤其有益于计算标准误差。第二种应用是，自助法能用作执行通常渐近理论的精炼，以此提供对检验统计量分布的更好的有限样本逼近。

在进入第 11 章完整研究之前，我们均用自助法实施沃尔德检验。

7.8.1 *利用标准渐近理论推断*

再次考虑 probit 例子，其中二值回归元 y 以概率 $p=\Phi(\gamma+\beta x)$ 等于 1，其中，$\Phi(\cdot)$ 表示标准正态 cdf。关注内容为，在显著性水平 0.05 上对 H_0：$\beta=1$ 与 H_a：$\beta\neq1$ 进行检验。这里的分析并不需要 probit 模型知识。

生成一个样本量 $N=40$ 的样本。使用 probit 极大似然估计，得出 $\hat{\beta}=0.817$ 且 $s_{\hat{\beta}}=0.294$，其中，标准误差建立在 $-\hat{\mathbf{A}}^{-1}$ 的基础上，因此，检验统计量 $z=(1-0.817)/0.294=-0.623$。

利用标准渐近理论，由于 $z_{0.25}=1.96$，所以得到 5% 的临界值为 -1.96 与 1.96，从而没有拒绝 H_0。

7.8.2 *不含渐近精炼的自助法*

自助法的出发点是从逼近再抽样到总体，参见 11.2.1 节。因此，通过从原始样本再抽样得到成对自助法。

因此，通过从初始数据 $\{(y_i,x_i),\ i=1,\cdots,N\}$ 中进行重复抽取，构成容量为 N 的 B 个拟样本。例如，第一个容量为 40 的拟样本，可以是 (y_1,x_1) 出现一次，而 (y_2,x_2) 出现两次，(y_3,x_3) 没有出现，等等。从而，得到关注参数 β 的 B 个估计值 $\hat{\beta}_1^*,\cdots,\hat{\beta}_B^*$，这些估计值用于估计初始 $\hat{\beta}$ 的分布。

例如，假定用于估计 probit 模型的计算机程序报告出 $\hat{\beta}$，但没有标准误差 $s_{\hat{\beta}}$。自助法可解决这一问题，因为能运用源于 B 个自助法拟样本的 $\hat{\beta}_1^*,\cdots,\hat{\beta}_B^*$ 估计标准差 $s_{\hat{\beta},\mathrm{boot}}$。当已知该标准误差估计时，可实施对 β 的沃尔德假设检验。

对 probit 沃尔德检验例子来说，所得到的 $\hat{\beta}$ 标准误差的自助法估计值是 0.376，进而得出 $z=(1-0.817)/0.376=-0.487$。由于 -0.487 位于 $(-1.96, 1.96)$ 中，故在 5%上不能拒绝 H_0。

用自助法进行检验假设，不会导致小样本水平的改进。不过，在许多应用中，如果用别的方法获得估计量的标准误差很困难，用这种方法能节省大量时间。

7.8.3 *带有渐近精炼的自助法*

某些自助法能使 z 分布具有更好的渐近逼近。在实际水平可能更接近于名义水平 0.05 的意义上，这样做可能获得更好的有限样本临界值。第 11 章对此给予详细讨论。现在，我们就阐述这一方法。

再次从初始数据中通过重复抽取得到拟容量为 N 的样本。在每个拟样本中估计 probit 模型，并对第 b 个拟样本计算 $z_b^*=(\hat{\beta}_b^*-\hat{\beta})/s_{\hat{\beta}_b^*}$，其中，$\hat{\beta}$ 表示初始估计值。于是，初始检验统计量 z 的自助法分布就是 $z_1^*,\cdots,z_B^*$ 的经验分布，而不是标准正态的。该经验分布的上 2.5 百分位数与下 2.5 百分位数给出了自助法临界值。

对上面例子来说，这里有 $B=1\,000$，求出 z 的经验自助法分布的上 2.5 百分位数与下 2.5 百分位数分别为 -1.89 与 1.80。于是，在 5%水平上进行检验，自助法临界值是 -2.62 与 1.83，而不是通常的 1.96。由于初始样本检验统计量 $z=-0.623$ 位于 $(-2.62, 1.83)$ 之内，故没有拒绝 H_0: $\beta=1$。同理，可计算自助法的 p 值。

和前一节的自助法不同，此处的渐近性得到了改进，这是因为学生化的检验统计量 z 是渐近中枢的（参见 11.2.3 节），而估计量 $\hat{\beta}$ 则不是。

7.9 应用研究

微观经济计量学考虑到要运用估计量的方差矩阵的稳健估计值，其重点为建立最小分布假设基础上的统计推断。然而，从稳健推断上来看，这样做没有什么意义，倘若分布假设失效，则会产生更严重的估计量非一致性的复杂性，尽管这种情况并不是对全部 MLS 估计量都会发生。

许多软件包在执行估计量命令时，都提供"稳健"标准误差选项。在微观经济计量学软件包中，稳健经常意味着异方差性一致，而并没有预防其他诸如聚集(clustering)的复杂性问题，参见 24.5 节，它还能引起无效的统计推断。

稳健推断通常利用沃尔德检验来实施。沃尔德检验对非线性假设的重新参数化来说，具有不变性弱点。尽管这可能通过执行适当的自助法来加以消除。虽然在一些情况下，LM 检验的相对简单稳健形式是可行的，但通常 LM 检验标准的辅助回归与 LM 检验的计算机软件包执行都不是稳健的（参见 8.4 节）。

检验的势可能是弱的。理想状态下，人们报告出对于某个有意义的备择假设的势。当这样做不行时，正如 7.6 节所述，人们应谨慎对待那些源自假设检验的结论，除非参数得到非常准确的估计。

此外,从渐近理论推导出的检验的有限样本本量是一个问题。第 11 章将详细阐述的自助法会潜在地得到假设检验与置信区间,并具有更好的有限样本性质。

统计推断可能是相当脆弱的,这些问题对实践者而言是重要的。当 $\hat{\theta}=1.96$ 时,考察统计显著性的双侧沃尔德检验,同时假定检验统计量实际上服从标准正态分布。如果 $s_{\hat{\theta}}=1.0$,那么 $t=1.96$ 且 p 值为 0.050。不过,当标准误差被低估 20%(因此,正确的 $t=1.57$)时,真实的 p 值就会很大,即 0.117;而当标准误差被高估 20%(因此,$t=2.35$)时,其真实 p 值是很小的,即 0.014。

7.10 文献注释

古里耶克斯和蒙福特(Gourieroux and Monfort, 1989)与戴维森和麦金农(Davidson and MacKinnon, 1993)所撰写的经济计量学教材,都对假设检验给予了详细阐述。本章阐述仅仅考察等式约束的情况。对于不等式约束的检验,参见古里耶克斯、霍利和蒙福特(Gouriéroux, Holly and Monfort, 1982)的线性情况,以及沃拉克(Wolak, 1991)的非线性情况。对假设检验来说,当在零假设下,参数位于参数空间的边界上时,检验就会失效,参见安德鲁斯(Andrews, 2001)。

7.3 三种经典检验方法的其中一种用图示论述,已由布斯(Buse, 1983)给出。

7.5 纽韦和韦斯特(Newey and West, 1987a)曾阐述经典检验对广义矩方法估计的扩展。

7.6 戴维森和麦金农(Davidson and MacKinnon, 1993)对势进行大量讨论,并解释显性零假设与备择假设和隐性零假设与备择假设之间的区别。

7.7 关于蒙特卡罗的研究,参见戴维森和麦金农(Davidson and MacKinnon, 1993),以及亨德里(Hendry, 1984)。

7.8 归功于埃弗龙(Efron, 1979)的自助法将在第 11 章详细阐释。

习 题

7-1 假定由一个样本得出估计值 $\hat{\theta}_1=5$,$\hat{\theta}_2=3$,其渐近方差估计值分别为 4 与 2,并且 $\hat{\theta}_1$ 与 $\hat{\theta}_2$ 的相关系数等于 0.5。假如参数估计值服从渐近正态性。

(a) 当水平为 0.05 时,H_0: $\theta_1 e^{\theta_2}=100$ 对 H_a: $\theta_1\neq 100$ 进行检验。

(b) 对于 $\gamma=\theta_1 e^{\theta_2}$,求 95%的置信区间。

7-2 考察模型 $y=\exp(\alpha+\beta x)+\varepsilon$ 的 NLS 回归,其中,α、β 以及 x 都表示纯量,且 $\varepsilon\sim\mathcal{N}[0,1]$。注意,为了简单起见,$\sigma_\varepsilon^2=1$,并不必估计。想要检验 H_0: $\beta=0$ 对 H_a: $\beta\neq 0$。

(a) 给出 α 与 β 的无约束 MLE 的一阶条件。

(b) 给出 α 与 β 的无约束渐近方差矩阵。

(c) 给出 α 与 β 的约束 MLE 的显式解。

(d) 为计算 LM 检验的 OPG 形式,请给出辅助回归。

(e) 对 LM 检验的初始形式,给出完整解释。注意,它将涉及在 α 与 β 的约束 MLE 处计算出的无约束对数似然的倒数。[这比(a)~(b)部分更困难。]

7-3 假定在两个嵌入式参数模型间进行选择。这两个模型密度的关系是 $g(y|x,\beta,\alpha=0)=f(y|x,\beta)$,为了简单起见,$\beta$ 与 α 都是纯量。如果 g 是正确的密度,那么建立在密度 f 基础上的 β 的 MLE 是非一致的。模型 f 对模型 g 进行检验,即 $H_0: \alpha=0$ 对 $H_a: \alpha\neq 0$ 进行检验。假定通过 ML 估计,得出下述结果。(1) 模型 f: $\hat{\beta}=5.0$, $\text{se}[\hat{\beta}]=0.5$ 以及 $\ln L=-106$;(2) 模型 g: $\hat{\beta}=3.0$, $\text{se}[\hat{\beta}]=1.0$, $\hat{\alpha}=2.5$, $\text{se}[\hat{\alpha}]=1.0$ 以及 $\ln L=-103$。已知前面信息,下述检验并不是全部可行的。倘若有足够信息,进行检验并叙述你的结论。若信息不够充分,请说明这一点。

(a) 在水平 0.05 上,实施 H_0 的沃尔德检验。

(b) 在水平 0.05 上,实施 H_0 的拉格朗日乘子检验。

(c) 在水平 0.05 上,实施 H_0 的似然比检验。

(d) 在水平 0.05 上,实施 H_0 的豪斯曼检验。

7-4 当数据生成过程为 $y\sim\mathcal{N}[\mu,100]$时,标准差为 10,并且样本量为 $N=10$,当名义水平为 0.05 时,考察 $H_0: \mu=0$ 对 $H_a: \mu\neq 0$ 的检验。检验统计量是通常的 t 检验统计量 $t=\hat{\mu}/\sqrt{s/10}$,其中,$s^2=(1/9)\sum_i(y_i-\bar{y})$。实施 10 000 次模拟,回答下述问题。

(a) 如果使用正确的有限样本临界值 $\pm t_{0.025}(8)=\pm 2.306$,求 t 检验的实际水平。存在水平扭曲吗?

(b) 如果使用渐近逼近临界值 $\pm z_{0.025}=\pm 1.960$,求 t 检验的实际水平。存在水平扭曲吗?

(c) 如果使用临界值 $\pm t_{0.025}(8)=\pm 2.306$,求 t 检验对备择假设 $H_a: \mu=1$ 的势。该检验对这个特定备择假设有势吗?

7-5 运用 16.6 节的健康支出数据。此模型是 DMED 的 probit 回归,DMED 表示良好健康支出的标示变量,对应的 17 个回归元已列在 16.6 节的第二段中。已知表 16.1 的第一列,你应求其估计值。在水平 0.05 上,考察自测健康[1](self-rated health)标示变量 HLTHG、HLTHF 以及 HLTHP 的统计显著性的联合检验。

(a) 实施沃尔德检验。

(b) 实施似然比检验。

(c) 为了执行 LM 检验,请提出一个辅助回归(这需要编写额外的某种程序)。

---

[1] self-rated health,中文译为自测健康,这个概念最早是由萨奇曼(Suchman)等人于 1958 年提出,它是个体对其健康状况的主观评价和期望。后来,许多学者对这一概念不断充实与完善。目前,自测健康法已成为国际上比较通用的健康测量方法之一。



设定检验与模型选择

8.1 引 论

在实际应用中，微观经济计量建模存在两个重要方面：一是确定模型是否被正确设定，二是对可供选择模型所进行的选取。就这两方面而言，尤其是当模型出现嵌套时，运用前一章阐述的假设检验方法是可行的。本章将阐述其他几种方法。

第一，m 检验，比如条件矩检验，是对模型所利用的矩条件是否得到满足而进行检验。除矩条件没有被利用到估计中而用于检验之外，这一方法在思想上类似广义矩方法(GMM)。这类检验在概念上与第 7 章假设的检验截然不同，因为可供选择的假设模型没有显式表述。

第二，豪斯曼检验是对两个估计量之间的差异进行检验，如果此模型被正确设定，那么两个估计量是一致的；但倘若模型被错误设定，则出现发散。

第三，对嵌套模型进行检验需要特殊方法，因为通常假设检验方法只有当一个模型嵌套在另一个模型之内时，才能应用。

最后，计算和报告那些作为非检验统计量的模型适合性统计量是有益的。例如，类似 R^2 形式可用于测算对非线性模型的拟合优度。

原则上讲，这些方法可用于模型设定、估计、检验和评价的全部过程。该整套过程能从一般模型到特殊模型，或者从特殊模型到更一般模型，用于捕获最重要的数据特征。

8.2 节阐述一些检验，包括条件矩检验、信息矩阵检验和卡方拟合优度检验。豪斯曼检验将在 8.3 节加以阐述。几种普遍的错误设定检验在 8.4 节讨论。8.5 节关注非嵌套式模型之间的区别。8.2 节～8.5 节普遍使用的容易执行的一些检验，都依赖于强分布，并且/或者在有限样本下执行效果不好。这种担心阻碍了对这些检验的部分运用，但此类担心已过时了，因为在许多情况下，将在第 11 章阐述的自助法可对这些弱点加以校正。8.6 节考虑对模型后续推断结果的检验。模型诊断将在 8.7 节加以阐述。

8.2 m 检验

m 检验，譬如条件矩检验，是一种一般的设定检验方法，它包括许多通行的设定检验。当使用极大似然法进行估计时，这种检验利用辅助回归就很容易执行，在这种情况下模型假设检验是尤其合适的。然而，当估计量建立在最小分布假设的基础上时，实施起来往往更困难一些。

首先，我们引进检验统计量与计算方法，然后通过重要例子阐明检验。

8.2.1 *m 检验*

假定模型蕴含总体矩条件(**population moment condition**)：

$$H_0:\ \mathrm{E}[\mathbf{m}_i(\mathbf{w}_i,\ \boldsymbol{\theta})]=\mathbf{0} \tag{8.1}$$

其中，$\mathbf{w}$ 表示可观测向量，通常因变量是 y，回归元是 $\mathbf{x}$，有时还有额外变量 $\mathbf{z}$，$\boldsymbol{\theta}$ 表示$q\times1$ 维参数向量，而 $\mathbf{m}_i(\cdot)$表示 $q\times1$ 维向量。当线性模型 $y=\mathbf{x}'\boldsymbol{\beta}+u$ 中的 $\mathbf{z}$ 被省略时，一个简单例子是 $\mathrm{E}[(y-\mathbf{x}'\boldsymbol{\beta})\mathbf{z}]=\mathbf{0}$。特别地，对完全参数模型来说，$\mathbf{m}_i(\cdot)$ 存在许多备选者。

m 检验(**m-test**)是对相应样本矩(**sample moment**)：

$$\hat{\mathbf{m}}_N(\hat{\boldsymbol{\theta}})=N^{-1}\sum_{i=1}^{N}\mathbf{m}_i(\mathbf{w}_i,\ \hat{\boldsymbol{\theta}}) \tag{8.2}$$

接近于 0 的检验，该方法类似于沃尔德检验，其中，$\mathbf{h}(\boldsymbol{\theta})=\mathbf{0}$ 表示对 $\mathbf{h}(\hat{\boldsymbol{\theta}})$ 接近于 0 所进行的检验。

检验统计量通过类似于 7.2.4 节所详述的沃尔德检验方法来获得。8.2.3 节将证明，若式(8.1)成立，则：

$$\sqrt{N}\hat{\mathbf{m}}_N(\hat{\boldsymbol{\theta}})\xrightarrow{d}\mathcal{N}[\mathbf{0},\ \mathbf{V}_{\mathbf{m}}] \tag{8.3}$$

其中，$\mathbf{V}_{\mathbf{m}}$ 由后面的式(8.10)定义，与沃尔德检验情况相比，$\mathbf{V}_{\mathbf{m}}$ 表现得更复杂，这是因为 $\mathbf{m}_i(\mathbf{w}_i,\ \hat{\boldsymbol{\theta}})$具有两个随机变异来源，此处的 $\mathbf{w}_i$ 和 $\hat{\boldsymbol{\theta}}$ 都是随机的。

于是，卡方检验统计量通过取相应的二次形式而得到。因此，式(8.1)的 m 检验统计量(**m-test statistic**)是：

$$\mathrm{M}=N\hat{\mathbf{m}}_N(\hat{\boldsymbol{\theta}})'\hat{\mathbf{V}}_{\mathbf{m}}^{-1}\hat{\mathbf{m}}_N(\hat{\boldsymbol{\theta}}) \tag{8.4}$$

若矩条件(8.1)成立，则 M 渐近服从 $\chi^2(\mathrm{rank}[\mathbf{V}_{\mathbf{m}}])$分布。当 $\mathrm{M}>\chi^2_\alpha(h)$时，在显著性水平 α 上，m 检验就拒绝矩条件(8.1)；否则，不拒绝矩条件(8.1)。

一种复杂情况是，$\mathbf{V}_{\mathbf{m}}$ 并不是满秩 h 的。例如，如果估计量 $\hat{\boldsymbol{\theta}}$ 本身设 $\hat{\mathbf{m}}_N(\hat{\boldsymbol{\theta}})$分量的线性组合为 $\mathbf{0}$，就是此种情况。在某些情况下，例如 OIR 检验，$\hat{\mathbf{V}}_{\mathbf{m}}$ 仍是满秩的，而且可计算出 M，但卡方检验统计量仅仅具有 $\mathrm{rank}[\mathbf{V}_{\mathbf{m}}]$个自由度。在另一些情况下，$\hat{\mathbf{V}}_{\mathbf{m}}$ 本身不是满秩的。于是，最简单的方法是删去$(h-\mathrm{rank}[\mathbf{V}_{\mathbf{m}}])$矩条件，并只利用这个矩条件的子集实施 m 检验。一种可供选择的方式是，使用全部矩条

件集合，只是式(8.4)中的 $\hat{\mathbf{V}}_{\mathrm{m}}^{-1}$ 要用 $\hat{\mathbf{V}}_{\mathrm{m}}^{-}$ 代替，$\hat{\mathbf{V}}_{\mathrm{m}}^{-}$ 表示 $\hat{\mathbf{V}}_{\mathrm{m}}$ 的广义逆。矩阵 $\mathbf{V}$ 的 Moore-Penrose 广义逆满足 $\mathbf{V}\mathbf{V}^{-}\mathbf{V}=\mathbf{V}$、$\mathbf{V}^{-}\mathbf{V}\mathbf{V}^{-}=\mathbf{V}^{-}$、$(\mathbf{V}\mathbf{V}^{-})'=\mathbf{V}\mathbf{V}^{-}$ 以及 $(\mathbf{V}^{-}\mathbf{V})'=\mathbf{V}^{-}\mathbf{V}$。当 $\mathbf{V}_{\mathrm{m}}$ 的秩比满秩小时，严格地讲，式(8.3)不再成立，因为多元正态要求满秩的 $\mathbf{V}_{\mathrm{m}}$，但给定这些条件，式(8.4)仍成立。

从概念上看，m 检验方法非常简单。当样本估计(8.2)的二次型离 0 甚远，就拒绝矩约束(8.1)。由于 $\hat{\mathbf{V}}_{\mathrm{m}}$ 是相当复杂的(参见 8.2.2 节)，并且需要选取矩 $\mathbf{m}(\cdot)$ 加以检验(参见 8.2.3 节～8.2.6 节的一些重要例子)和解释拒绝式(8.1)的理由(参见 8.2.2 节)，因此，计算 M 是一个挑战。

8.2.2 *计算 m 统计量*

存在几种计算 m 统计量的方法。

第一，利用 8.2.3 节给出的 $\mathbf{V}_{\mathrm{m}}$ 分量一致估计，总是可以直接计算 $\hat{\mathbf{V}}_{\mathrm{m}}$，从而计算 M。大多数应用研究者都避开这种方法，因为它需要矩阵计算。

第二，运用自助法(**bootstrap**)(参见 11.6.3 节)，因为自助法能提供控制 $\hat{\mathbf{m}}_N(\hat{\boldsymbol{\theta}})=N^{-1}\sum_i\mathbf{m}_i(\mathbf{w}_i,\hat{\boldsymbol{\theta}})$ 中所有变异来源的 $\mathbf{V}_{\mathrm{m}}$ 估计值。

第三，在某些情况下，类似于 7.3.5 节给出的 LM 检验情况，运用辅助回归(**auxiliary regressions**)能计算 M 的渐近等价形式，而这并不需要计算 $\hat{\mathbf{V}}_{\mathrm{m}}$。这些辅助回归也可利用自助法，以便获得渐近精炼(参见 11.6.3 节)。我们将阐述几种重要的辅助回归。

利用极大似然估计量的辅助回归

当在似然框架下进行推断时，模型设定检验尤其是值得做的，因为通常对密度的任何错误设定，都能导致极大似然估计的非一致性。幸运的是，当运用极大似然估计时，容易实施 m 检验。

具体地讲，当 $\hat{\boldsymbol{\theta}}$ 是极大似然估计值时，7.3.5 节推广的 LM 检验结果会产生下述情况：m 检验的渐近等价形式由辅助回归(**auxiliary regression**)

$$1=\hat{\mathbf{m}}_i'\boldsymbol{\delta}+\hat{\mathbf{s}}_i'\boldsymbol{\gamma}+u_i \tag{8.5}$$

获得，其中，$\hat{\mathbf{m}}_i=\mathbf{m}_i(y_i,\mathbf{x}_i,\hat{\boldsymbol{\theta}}_{\mathrm{ML}})$，$\hat{\mathbf{s}}_i=\partial\ln f(y_i|\mathbf{x}_i,\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\hat{\theta}_{\mathrm{ML}}}$ 表示第 i 个观测值对得分的贡献，而 $f(y_i|\mathbf{x}_i,\boldsymbol{\theta})$ 表示条件密度函数，这通过

$$\mathrm{M}^*=NR_u^2 \tag{8.6}$$

来计算，其中，R_u^2 表示在 7.3.5 节结尾处定义的非中心 R^2。等价地讲，M^* 等于 ESS_u，即源自回归(8.5)的非中心解释平方和(拟合值平方和)；或者，M^* 等于 $N-\mathrm{RSS}$，这里的 RSS 表示源自回归(8.5)的残差平方和。在 H_0 下，M^* 渐近服从 $\chi^2(h)$。

检验统计量 M^* 被称为 m 检验的梯度外积(**outer product of the gradient**，简记为 OPG)形式，并且它是 LM 检验辅助回归的推广(参见 7.3.5 节)。尽管容易计算外积梯度形式，但它具有大水平扭曲(**large size distortions**)不好的小样本性质。但是，类似于 LM 检验，这些小样本问题可以很方便地利用自助法来加以减少(参

见 11.6.3 节)。

在某些非极大似然背景下,检验统计量 M^*同样是适宜的。每当 $\mathrm{E}[\partial \mathbf{m}/\partial \boldsymbol{\theta}']=-\mathrm{E}[\mathbf{ms}']$,就可应用辅助回归(参见 8.2.3 节)。由广义信息矩阵等式(参见 5.6.3 节),对极大似然估计来说,当期望是设定密度函数 $f(\cdot)$时,这个条件就成立。在一些情况下,在比较弱分布的假设下,它同样是成立的。

当 $\mathrm{E}[\partial \mathbf{m}/\partial \boldsymbol{\theta}']=\mathbf{0}$ 时的辅助回归

在一些应用中,除满足式(8.1)之外,$\mathbf{m}_i(\mathbf{w}_i, \boldsymbol{\theta})$满足:

$$\mathrm{E}[\partial \mathbf{m}_i(\mathbf{w}_i, \boldsymbol{\theta})/\partial \boldsymbol{\theta}'|_{\theta_0}]=\mathbf{0} \tag{8.7}$$

于是,可以证明 $\sqrt{N}\hat{\mathbf{m}}_N(\hat{\boldsymbol{\theta}})$的渐近分布与 $\sqrt{N}\mathbf{m}_N(\boldsymbol{\theta}_0)$的一样,因此,$\mathbf{V}_m=\operatorname{plim} N^{-1}\sum_i \mathbf{m}_{i0}\mathbf{m}_{i0}'$,这可通过 $\hat{\mathbf{V}}_m=N^{-1}\sum_i \hat{\mathbf{m}}_i\hat{\mathbf{m}}_i'$一致估计出。除辅助回归(**auxiliary regression**)是更简单的

$$1=\hat{\mathbf{m}}_i'\boldsymbol{\delta}+u_i \tag{8.8}$$

之外,此检验统计量可类似于式(8.5)加以计算,检验统计量 M^{**} 等于 N 倍的非中心化R^2。

倘若式(8.7)成立,不像极大似然估计那样,对任何根号 N 的一致估计量$\hat{\boldsymbol{\theta}}$来说,这个辅助回归都是有效的。少数例子均会遇到条件(8.7);参见 8.2.9 节的例子。

即使式(8.7)不成立,比较简单回归(8.8)仍然可以作为一个指南,这是因为它对 M 的正确值即 m 检验统计量施加了一个下界。当拒绝这个较简单回归时,就一定拒绝式(8.1)。

其他辅助回归

若 $\mathbf{m}(y,\mathbf{x},\boldsymbol{\theta})$与 $\mathbf{s}(y,\mathbf{x},\boldsymbol{\theta})$能适当地因式分解,则对式(8.5)与式(8.8)进行可供选择的辅助回归是可能的。

第一,对某些满足 $\mathrm{V}[r(y,\mathbf{x},\boldsymbol{\theta})]=1$ 的共同纯量函数 $r(\cdot)$来说,若 $\mathbf{s}(y,\mathbf{x},\boldsymbol{\theta})=\mathbf{g}(\mathbf{x},\boldsymbol{\theta})r(y,\mathbf{x},\boldsymbol{\theta})$且 $\mathbf{m}(y,\mathbf{x},\boldsymbol{\theta})=\mathbf{h}(\mathbf{x},\boldsymbol{\theta})r(y,\mathbf{x},\boldsymbol{\theta})$,通过极大似然法估计,渐近等价于式(8.5)的回归就是源于 $\hat{r}_i$ 对 $\hat{\mathbf{g}}_i$ 与 $\hat{\mathbf{h}}_i$ 回归的 NR_u^2。

第二,对某个满足 $\mathrm{V}[v(y,\mathbf{x},\boldsymbol{\theta})]=1$ 的纯量函数 $v(\cdot)$来说,若 $\mathbf{m}(y,\mathbf{x},\boldsymbol{\theta})=\mathbf{h}(\mathbf{x},\boldsymbol{\theta})v(y,\mathbf{x},\boldsymbol{\theta})$且 $\mathrm{E}[\partial \mathbf{m}/\partial \boldsymbol{\theta}']=\mathbf{0}$,则渐近等价于式(8.8)的回归是源于 $\hat{v}_i$ 对 $\hat{\mathbf{h}}_i$ 回归的 NR_u^2。有关更详细的内容,参见伍德里奇(Wooldridge, 1991)。

在特殊背景下,存在另一些辅助回归。8.4 节将给出一些例子,怀特(White, 1994)对此给出相当一般的研究。

8.2.3 *m 检验统计量的推导*

为了避免计算 $\mathbf{V}_m$ 即式(8.3)中的方差矩阵,m 检验通常利用辅助回归或自助法来实施。为了完整起见,本节将推导 $\mathbf{V}_m$ 的实际表达式,同时提供判断辅助回归(8.5)与式(8.8)的正确理由。

一个关键内容是获得式(8.2)定义的 $\hat{\mathbf{m}}_N(\hat{\boldsymbol{\theta}})$分布。想要得到前面的分布极为

复杂，因为 $\mathbf{m}_N(\hat{\boldsymbol{\theta}})$ 是随机的，其原因有两个：一个是随机变量 $\mathbf{w}_i$，另一个是在估计量 $\hat{\boldsymbol{\theta}}$ 处计算。

假定 $\hat{\boldsymbol{\theta}}$ 是 m 估计量或估计方程估计量，对于某个函数 $\mathbf{s}(\cdot)$，它是：

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{s}_i(\mathbf{w}_i,\hat{\boldsymbol{\theta}})=\mathbf{0} \tag{8.9}$$

的解，这里不一定有 $\partial\ln f(y|\mathbf{x},\boldsymbol{\theta})/\partial\boldsymbol{\theta}$，并做出通常横截面假设：对于不同 i，数据是独立的。从而，可以证明 $\sqrt{N}\hat{\mathbf{m}}_N(\hat{\boldsymbol{\theta}})\xrightarrow{d}\mathcal{N}[\mathbf{0},\mathbf{V_m}]$，如同式(8.3)一样，其中：

$$\mathbf{V_m}=\mathbf{H}_0\mathbf{J}_0\mathbf{H}_0' \tag{8.10}$$

$\mathbf{H}_0$ 为一个 $h\times(h+q)$ 阶矩阵，即：

$$\mathbf{H}_0=[\mathbf{I}_h\ -\mathbf{C}_0\mathbf{A}_0^{-1}] \tag{8.11}$$

其中，$\mathbf{C}_0=\text{plim}\,N^{-1}\sum_i\partial\mathbf{m}_{i0}/\partial\boldsymbol{\theta}'$，$\mathbf{A}_0=\text{plim}\,N^{-1}\sum_i\partial\mathbf{s}_{i0}/\partial\boldsymbol{\theta}'$，并且有 $(h+q)(h+q)$ 阶矩阵：

$$\mathbf{J}_0=\text{plim}\,N^{-1}\begin{bmatrix}\sum_{i=1}^{N}\mathbf{m}_{i0}\mathbf{m}_{i0}' & \sum_{i=1}^{N}\mathbf{m}_{i0}\mathbf{s}_{i0}'\\ \sum_{i=1}^{N}\mathbf{s}_{i0}\mathbf{m}_{i0}' & \sum_{i=1}^{N}\mathbf{s}_{i0}\mathbf{s}_{i0}'\end{bmatrix} \tag{8.12}$$

其中，$\mathbf{m}_{i0}=\mathbf{m}_i(\mathbf{w}_i,\boldsymbol{\theta}_0)$，$\mathbf{s}_{i0}=\mathbf{s}_i(\mathbf{w}_i,\boldsymbol{\theta}_0)$。

为了推导式(8.10)，在 $\boldsymbol{\theta}_0$ 附近实施一阶泰勒级数展开，得到：

$$\sqrt{N}\hat{\mathbf{m}}_N(\hat{\boldsymbol{\theta}})=\sqrt{N}\mathbf{m}_N(\boldsymbol{\theta}_0)+\frac{\partial\mathbf{m}_N(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}'}\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)+o_p(1) \tag{8.13}$$

对于式(8.9)定义的 $\hat{\boldsymbol{\theta}}$ 来说，这蕴含：

$$\sqrt{N}\hat{\mathbf{m}}_N(\hat{\boldsymbol{\theta}})=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{m}_i(\boldsymbol{\theta}_0)-\mathbf{C}_0\mathbf{A}_0^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{s}_{i0}+o_p(1) \tag{8.14}$$

其中，使用了 $\mathbf{m}_N=N^{-1}\sum_i\mathbf{m}_i$，$\partial\mathbf{m}_N/\boldsymbol{\theta}'=N^{-1}\sum_i\partial\mathbf{m}_i/\partial\boldsymbol{\theta}'\xrightarrow{p}\mathbf{C}_0$，并且把通常的一阶泰勒级数展开式用于式(8.9)，$\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$ 与 $\mathbf{A}_0^{-1}N^{-1/2}\sum_i\mathbf{s}_{i0}$ 具有相同的极限分布。将式(8.14)写成：

$$\sqrt{N}\hat{\mathbf{m}}_N(\hat{\boldsymbol{\theta}})=[\mathbf{I}_h\quad -\mathbf{C}_0\mathbf{A}_0^{-1}]\begin{bmatrix}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{m}_{i0}\\ \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{s}_{i0}\end{bmatrix}+o_p(1) \tag{8.15}$$

通过运用极限正态积准则(定理 A.17)，得出式(8.10)，这是因为式(8.15)中积的第二项在 H_0 下服从极限正态分布，其均值为 $\mathbf{0}$，方差为 $\mathbf{J}_0$。

为了计算式(8.4)中的 M，通过使用一致估计值代替 $\mathbf{V_m}$ 的每一个分量，得出 $\mathbf{V_m}$ 的一致估计值 $\hat{\mathbf{V}}_m$。例如，通过 $\hat{\mathbf{C}}=N^{-1}\sum_i\partial\mathbf{m}_i/\partial\boldsymbol{\theta}'|_{\hat{\theta}}$ 来一致估计 $\mathbf{C}_0$，等等。这样做尽管总是可行的，但当利用辅助回归时就比较容易。

第一，当 $\hat{\boldsymbol{\theta}}$ 是极大似然估计值时，考察辅助回归(8.5)。由广义信息矩阵等式(参见5.6.3节)，$\mathrm{E}[\partial \mathbf{m}_{i0}/\partial \boldsymbol{\theta}']=-\mathrm{E}[\mathbf{m}_{i0}\mathbf{s}_{i0}']$，其中，对极大似然估计值来说，我们专门研究 $\mathbf{s}_i=\partial \ln f(y_i,\mathbf{x}_i,\boldsymbol{\theta})/\partial \boldsymbol{\theta}'$。由于可进行大量简化，$\mathbf{C}_0=-\operatorname{plim} N^{-1}\times\sum_i \mathbf{m}_{i0}\mathbf{s}_{i0}'$，而 $\mathbf{A}_0=-\operatorname{plim} N^{-1}\sum_i \mathbf{s}_{i0}\mathbf{s}_{i0}'$，这也出现在 $\mathbf{J}_0$ 矩阵中。故得到检验的OPG形式。有关更详细的内容，参见纽韦(Newey, 1985)，以及帕甘和维拉(Pagan and Vella, 1989)。

第二，就辅助回归(8.8)而言，注意到，若 $\mathrm{E}[\partial \mathbf{m}_{i0}/\partial \boldsymbol{\theta}']=\mathbf{0}$，则 $\mathbf{C}_0=\mathbf{0}$，因此 $\mathbf{H}_0=[\mathbf{I}_h \quad \mathbf{0}]$，从而 $\mathbf{H}_0\mathbf{J}_0\mathbf{H}_0'=\operatorname{plim} N^{-1}\sum_i \mathbf{m}_{i0}\mathbf{m}_{i0}'$。

8.2.4 *条件矩检验*

归功于纽韦(Newey, 1985)和陶亨(Tauchen, 1985)的条件矩检验，是对无条件矩约束的m检验，该无条件矩约束由基本条件矩约束来获得。

举一个例子，考察线性回归模型 $y=\mathbf{x}'\boldsymbol{\beta}+u$。关于OLS估计量一致性的标准假设是，误差具有条件零均值，或等价地，为条件矩约束：

$$\mathrm{E}[y-\mathbf{x}'\boldsymbol{\beta} \mid \mathbf{x}]=\mathbf{0} \tag{8.16}$$

第6章曾考察利用某些隐含无条件矩约束作为矩方法或广义矩方法估计的基础。特别地，式(8.16)蕴含 $\mathrm{E}[\mathbf{x}(y-\mathbf{x}'\boldsymbol{\beta})]=\mathbf{0}$。求解相应的样本矩条件 $\sum_i \mathbf{x}_i(y_i-\mathbf{x}_i'\boldsymbol{\beta})=\mathbf{0}$，得出 $\boldsymbol{\beta}$ 的OLS估计量。不过，式(8.16)蕴含，许多其他矩条件在估计中没有得到应用。考察无条件矩约束：

$$\mathrm{E}[\mathbf{g}(\mathbf{x})(y-x'\boldsymbol{\beta})]=\mathbf{0}$$

其中，向量 $\mathbf{g}(\mathbf{x})$ 应该不同于 $\mathbf{x}$，这已在OLS估计中使用过。例如，$\mathbf{g}(\mathbf{x})$ 可以包括回归元向量 $\mathbf{x}$ 分量的平方项或者交叉积。这表明，建立在相应样本矩 $\hat{\mathbf{m}}_N(\hat{\boldsymbol{\beta}})=N^{-1}\sum_i \mathbf{g}(\mathbf{x}_i)(y_i-\mathbf{x}_i'\hat{\boldsymbol{\beta}})$ 基础上的检验是否接近于0。

更一般地讲，对某一个纯量函数 $r(\cdot)$，考察条件矩约束：

$$\mathrm{E}[r(y,\mathbf{x},\boldsymbol{\theta}) \mid \mathbf{x}]=\mathbf{0} \tag{8.17}$$

条件矩检验[**conditional (CM) moment test**]是建立在隐含无条件矩约束

$$\mathrm{E}[\mathbf{g}(\mathbf{x})r(y,\mathbf{x},\boldsymbol{\theta})]=\mathbf{0} \tag{8.18}$$

基础上的m检验，其中，对 $\mathbf{g}(\mathbf{x})$ 与/或 $r(y,\mathbf{x},\boldsymbol{\theta})$ 进行选取，以使这些约束没有用于估计之中。

基于似然模型会导致许多潜在约束。比 $r(y,\mathbf{x},\boldsymbol{\theta})$ 完全参数模型稍差一些的例子包括 $y-\mu(\mathbf{x},\boldsymbol{\theta})$，以及 $(y-\mu(\mathbf{x},\boldsymbol{\theta}))^2-\sigma^2(\mathbf{x},\boldsymbol{\theta})$，其中，$\mu(\cdot)$ 表示设定条件均值函数，$\sigma^2(\mathbf{x},\boldsymbol{\theta})$ 表示设定条件方差函数。

8.2.5 *怀特信息矩阵检验*

对极大似然估计来说，信息矩阵等式蕴含可能用于m检验的一些矩约束，因为它们通常在求极大似然估计值时没有得到利用。

具体地讲，由 5.6.3 节的信息矩阵等式得出：

$$\mathrm{E}[\mathrm{Vech}[\mathbf{D}_i(y_i,\mathbf{x}_i,\boldsymbol{\theta}_0)]]=\mathbf{0} \tag{8.19}$$

其中，$q\times q$ 阶矩阵 $\mathbf{D}_i$ 由

$$\mathbf{D}_i(y_i,\mathbf{x}_i,\boldsymbol{\theta}_0)=\frac{\partial^2\ln f_i}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}+\frac{\partial\ln f_i}{\partial\boldsymbol{\theta}}\frac{\partial\ln f_i}{\partial\boldsymbol{\theta}'} \tag{8.20}$$

给出，而期望是针对假定的条件密度 $f_i=f(y_i|\mathbf{x}_i,\boldsymbol{\theta})$ 选取的。这里，Vech 表示半向量算子(**vector-half operator**)，以与向量算子(**Vec operator**)相同的方式对矩阵 $\mathbf{D}_i$ 的列进行叠放，只是对称矩阵 $\mathbf{D}_i$ 的仅仅 $q(q+1)/2$ 个元素得以叠放。

怀特(White, 1982)提出了相应的样本矩

$$\hat{\mathbf{d}}_N(\hat{\boldsymbol{\theta}})=N^{-1}\sum_{i=1}^{N}\mathrm{Vech}[\mathbf{D}_i(y_i,\mathbf{x}_i,\hat{\boldsymbol{\theta}}_{\mathrm{ML}})] \tag{8.21}$$

是否接近于 0 的信息矩阵检验(**information matrix test**)。利用式(8.4)，信息矩阵检验统计量是：

$$\mathrm{IM}=N\hat{\mathbf{d}}_N(\hat{\boldsymbol{\theta}})'\hat{\mathbf{V}}^{-1}\hat{\mathbf{d}}_N(\hat{\boldsymbol{\theta}}) \tag{8.22}$$

其中，怀特(White, 1982)曾给出的关于 $\hat{\mathbf{V}}$ 的表达式是相当复杂的。归功于兰开斯特(Lancaster, 1984)与切舍(Chesher, 1984)的更加容易实施的此检验方法，是运用辅助回归(8.5)，由于 MLE 可用于式(8.21)，所以辅助回归(8.5)是可应用的。

信息矩阵检验还能用于式(8.19)的约束中的子集上。确实应该这样做，当 q 很大时，进而用于检验的约束个数 $q(q+1)/2$ 就非常大。

当信息矩阵检验统计量的值很大，则拒绝信息矩阵等式约束，并得出密度被错误设定的结论。通常，这意味着极大似然估计量是非一致的。在一些特殊情况下，尽管标准误差需要建立在方差矩阵三明治形式的基础上，5.7 节已经详述，极大似然估计还是一致的。

8.2.6 *卡方拟合优度检验*

对完全参数模型来说，一个有用的设定检验是把预测概率与样本有关频率进行比较。当这些比较相差甚远，该模型就不是一个好模型。

以离散 iid 随机变量 y 来开始，y 以概率 $p_1,p_2,\cdots,p_J$ 取 J 个可能值之一，$\sum_{j=1}^{J}p_j=1$。对概率正确设定，可通过对理论上的频率 Np_j 等于观测频率 $N\bar{p}_j$ 这一等式进行检验来加以确定，其中，$\bar{p}_j$ 表示样本取第 j 个可能值的小数。皮尔逊卡方拟合优度检验统计量[**Pearson chi-square goodness-of-fit text (PCGF) statistic**]是：

$$\mathrm{PCGF}=\sum_{j=1}^{J}\frac{(N\bar{p}_j-Np_j)^2}{Np_j} \tag{8.23}$$

在零假设：概率 $p_1,p_2,\cdots,p_J$ 是正确的条件下，该统计量渐近服从 $\chi^2(J-1)$ 分布。对此检验加以推广，以便利用回归预测其概率(参见习题 8.2)。考察离散 y 具有

概率 $p_{ij}=p_{ij}(\mathbf{x}_i,\boldsymbol{\theta})$的多项式模型。于是,用 $\hat{p}_j=N^{-1}\sum_i F_j(\mathbf{x}_i,\hat{\boldsymbol{\theta}})$代替式(8.23)中的 p_j,而且如果 $\hat{\boldsymbol{\theta}}$ 是多项式 MLE,那么再次得到卡方分布,只是因估计 $\boldsymbol{\theta}$ 而减少自由度个数($J-\dim(\boldsymbol{\theta})-1$)[参见安德鲁斯(Andrews,1988a)]。

除多项式模型之外,对回归模型来说,式(8.23)的统计量 PCGF 可通过把 y 分成胞腔加以计算,但统计量 PCGF 已不再服从卡方分布。不过,可使用密切相关的 m 检验。为了推导这一统计量,把 y 的范围分割成 J 个互斥胞腔,这 J 个胞腔张成了 y 的所有可能值。设 $d_{ij}(y_i)$表示标示变量,当 y_i 属于第 j 个胞腔时,则它等于1,否则等于0。设 $p_{ij}(\mathbf{x}_i,\boldsymbol{\theta})=\int_{y_i\in 第j个胞腔} f(y_i\mid \mathbf{x}_i,\boldsymbol{\theta})\,dy_i$ 表示第 i 个观测值落入第 j 个胞腔的预测概率,其中,$f(y|\mathbf{x},\boldsymbol{\theta})$表示 y 的条件密度,同时首先假定参数向量 $\boldsymbol{\theta}$ 是已知的。若条件密度被正确设定,则:

$$\mathrm{E}[d_{ij}(y_i)-p_{ij}(\mathbf{x}_i,\boldsymbol{\theta})]=0,\quad j=1,\cdots,J \tag{8.24}$$

一旦以明确向量记号表示叠放所有 J 个矩,得出:

$$\mathrm{E}[\mathbf{d}_i(y_i)-\mathbf{p}_i(\mathbf{x}_i,\boldsymbol{\theta})]=\mathbf{0} \tag{8.25}$$

其中,$\mathbf{d}_i$ 与 $\mathbf{p}_i$ 均表示 $J\times 1$ 维向量,其第 j 个元素分别为 d_{ij} 与 p_{ij}。这就建议相应样本矩接近于0的 m 检验:

$$\widehat{\mathbf{dp}}_N(\hat{\boldsymbol{\theta}})=N^{-1}\sum_{i=1}^{N}(\mathbf{d}_i(y_i)-\mathbf{p}_i(\mathbf{x}_i,\hat{\boldsymbol{\theta}})) \tag{8.26}$$

它是样本有关频率向量 $N^{-1}\sum_i\mathbf{d}_i$ 与预测频率向量 $N^{-1}\sum_i\hat{\mathbf{p}}_i$ 之差。利用式(8.5),就得出安德鲁斯(Andrews, 1988a, 1988b)的卡方拟合优度检验统计量[**chi-square goodness-of fit (CGF) test statistic**]:

$$\mathrm{CGF}=N\,\widehat{\mathbf{dp}}_N(\hat{\boldsymbol{\theta}})'\hat{\mathbf{V}}^{-1}\widehat{\mathbf{dp}}_N(\hat{\boldsymbol{\theta}}) \tag{8.27}$$

其中,$\hat{\mathbf{V}}$ 的表达式是相当复杂的。利用辅助回归(8.5)以及 $\hat{\mathbf{m}}_i=\mathbf{d}_i-\hat{\mathbf{p}}_i$,容易计算 CGF 检验统计量。这个辅助回归是适宜的,因为完全参数模型得到了检验,从而 $\hat{\boldsymbol{\theta}}$ 是 MLE。

在 $f(y|\mathbf{x},\boldsymbol{\theta})$被正确设定的假设下,所得到的检验统计量渐近服从 $\chi^2(J-1)$,由于概率之和为1是一个约束,所以需要去掉一个分类。进一步地,在一些特殊情况下,可能要去掉一些分类,譬如在式(8.23)后面曾讨论的多项式例子。除报告已计算的检验统计量之外,报告 $N^{-1}\sum_i\mathbf{d}_i$ 与 $N^{-1}\sum_i\hat{\mathbf{p}}_i$ 的分量是有价值的。

安德鲁斯(Andrews, 1988a, 1988b)已提供有关的渐近理论,他给出比较简单的表述和几个应用。为了简单起见,我们依据 y 的范围来决定所述胞腔,不过这种划分既可依据 y 又可依据 $\mathbf{x}$ 而定。应该对胞腔进行选取,以便不存在仅有几个观测值的胞腔。对于更详细的内容和这种检验的历史,参见安德鲁斯的这些论文。

在连续随机变量 y 为 iid 的情况下,比 SCGF 检验更为一般的检验是柯尔莫哥洛夫检验(**Kolmogorov test**);此检验使用 y 的整个分布,而不是由 y 胞腔所形成的分布。安德鲁斯(Andrews, 1997)曾经阐述柯尔莫哥洛夫检验的回归形式,但是,

与卡方拟合优度检验相比，它却显得更加难以实施。

8.2.7 过度识别约束检验

对过度识别假设进行检验（参见 6.3.8 节）是 m 检验的一个例子。

运用第 6 章的记号，广义矩方法估计量是建立在 $\mathrm{E}[\mathbf{h}(\mathbf{w}_i,\boldsymbol{\theta}_0)]=\mathbf{0}$ 假设的基础上。若模型是过度识别的，则这些矩约束中仅有 q 个用于估计，从而得出 $(r-q)$ 个线性相关正交性条件，其中，$r=\dim[\mathbf{h}(\cdot)]$，这能用于构建 m 检验。于是，我们使用式（8.4）中的 M，其中，$\hat{\mathbf{m}}_N=N^{-1}\sum_i\mathbf{h}(\mathbf{w}_i,\hat{\boldsymbol{\theta}})$。正如 6.3.9 节表明的，若 $\hat{\boldsymbol{\theta}}$ 是最优广义矩方法估计量，则 $\hat{\mathbf{m}}_N(\hat{\boldsymbol{\theta}})'\hat{\mathbf{S}}_N^{-1}\hat{\mathbf{m}}_N(\hat{\boldsymbol{\theta}})$ 渐近服从 $\chi^2(r-q)$ 分布，其中，$\hat{\mathbf{S}}_N=N^{-1}\sum_{i=1}^N\hat{\mathbf{h}}_i\hat{\mathbf{h}}_i'$。更直观的线性工具变量例子将由 8.4.4 节给出。

8.2.8 条件矩检验的势与一致性

由于不存在显性可供选择的假设，所以 m 检验不同于第 7 章的检验。

有几位作者已经给出一些例子，可以证明，例子中的 IM 检验等价于传统的零假设对备择假设的 LM 检验。切舍（Chesher，1984）把 IM 检验解释成为对随机参数异质性的检验。对于正态性条件下的线性模型来说，霍尔（Hall，1987）已经证明，IM 检验的子分类对应于异方差性、对称性以及峰度的 LM 检验。卡梅伦和特里维迪（Cameron and Trivedi，1998）已给出线性指数族结果的某些其他例子和参考文献。

更一般地讲，m 检验能在下述条件矩框架下加以解释。以线性回归模型中对添加变量进行检验来开始。假定想要检验模型 $y=\mathbf{x}_1'\boldsymbol{\beta}_1+\mathbf{x}_2'\boldsymbol{\beta}_2+u$ 中是否有 $\boldsymbol{\beta}_2=\mathbf{0}$。这是 H_0：$\mathrm{E}[y-\mathbf{x}_1'\boldsymbol{\beta}_1|\mathbf{x}]=0$ 对 H_a：$\mathrm{E}[y-\mathbf{x}_1'\boldsymbol{\beta}_1|\mathbf{x}]=\mathbf{x}_2'\boldsymbol{\beta}_2$ 的检验。在 $y-\mathbf{x}_1'\boldsymbol{\beta}_1$ 对 $\mathbf{x}_2$ 的回归中，在 H_0 下并且假定对于不同 i 具有独立性，H_a：$\boldsymbol{\beta}_2=\mathbf{0}$ 的最强有力的检验是建立在有效 GLS 估计量

$$\hat{\boldsymbol{\beta}}_2=\left[\sum_{i=1}^N\frac{\mathbf{x}_{2i}\mathbf{x}_{2i}'}{\sigma_i^2}\right]^{-1}\sum_{i=1}^N\frac{\mathbf{x}_{2i}(y_i-\mathbf{x}_{1i}'\boldsymbol{\beta}_1)}{\sigma_i^2}$$

的基础上，其中，$\sigma_i^2=\mathrm{V}[y_i|\mathbf{x}_i]$，该检验等价于仅仅建立在第二个和式基础上的检验，它是

$$\mathrm{E}\left[\frac{\mathbf{x}_{2i}(y_i-\mathbf{x}_{1i}'\boldsymbol{\beta}_1)}{\sigma_i^2}\right]=\mathbf{0}\tag{8.28}$$

的 m 检验。一旦对过程加以颠倒，就把建立在式（8.28）基础上的 m 检验解释成为 H_0：$\mathrm{E}[y-\mathbf{x}_1'\boldsymbol{\beta}_1|\mathbf{x}]=0$ 对 H_a：$\mathrm{E}[y-\mathbf{x}_1'\boldsymbol{\beta}_1|\mathbf{x}]=\mathbf{x}_2'\boldsymbol{\beta}_2$ 的 CM 检验。同理，把建立在 $\mathrm{E}[\mathbf{x}_2(y-\mathbf{x}_1'\boldsymbol{\beta}_1)]=\mathbf{0}$ 基础上的 m 检验解释成为 H_0：$\mathrm{E}[y-\mathbf{x}_1'\boldsymbol{\beta}_1|\mathbf{x}]=0$ 对 H_a：$\mathrm{E}[y-\mathbf{x}_1'\boldsymbol{\beta}_1|\mathbf{x}]=\sigma_{y|\mathbf{x}}^2\mathbf{x}_2'\boldsymbol{\beta}_2$，其中，在 H_0 下，有 $\sigma_{y|\mathbf{x}}^2=\mathrm{V}[y|\mathbf{x}]$。

更一般地讲，对于某个纯量函数 $r(\cdot)$，假定以条件矩约束

$$\mathrm{E}[r(y_i,\mathbf{x}_i,\boldsymbol{\theta})\ |\mathbf{x}_i]=0\tag{8.29}$$

开始。于是，建立在无条件矩约束

$$\mathrm{E}[\mathbf{g}(\mathbf{x}_i)r(y_i,\mathbf{x}_i,\boldsymbol{\theta})]=\mathbf{0} \tag{8.30}$$

基础上的 m 检验,可能被解释成具有下述零假设与备择假设的 CM 检验:

$$H_0: \mathrm{E}[r(y_i,\mathbf{x}_i,\boldsymbol{\theta})|\mathbf{x}_i]=0 \tag{8.31}$$
$$H_a: \mathrm{E}[r(y_i,\mathbf{x}_i,\boldsymbol{\theta})|\mathbf{x}_i]=\sigma_i^2\mathbf{g}(\mathbf{x}_i)'\boldsymbol{\gamma}$$

其中,在 H_0 下,有 $\sigma_i^2=\mathrm{V}[r(y_i,\mathbf{x}_i,\boldsymbol{\theta})|\mathbf{x}_i]$。

这个方法给出了在哪个方向上 CM 检验具有势的指南。尽管式(8.30)表明,势通常位于 $\mathbf{g}(\mathbf{x})$的方向上,由式(8.31)知,更准确的表述是,用 $r(y,\mathbf{x},\boldsymbol{\theta})$方差乘以 $\mathbf{g}(\mathbf{x})$的方向。这个差异是重要的,因为在许多横截面应用中,对不同观测值而言,这个方差不为常值。对于更详细的内容及参考文献,可参见卡梅伦和特里维迪(Cameron and Trivedi, 1998),他们称这为基于回归的 CM 检验。尽管此方法表现出更繁琐的代数运算,但可把它推广到向量 $\mathbf{r}(\cdot)$上。

m 检验是对有限多个矩约束所进行的检验。因此,对基本条件的矩条件进行数据生成过程是可行的,例如,式(8.29)中的条件矩约束不正确,但矩条件却得到满足。于是,CM 检验是非一致的,当 $N\to\infty$时,以概率 1 不能拒绝。为了对非线性回归模型[其中,$r(y,\mathbf{x},\boldsymbol{\theta})=y-f(\mathbf{x},\boldsymbol{\theta})$]中的函数形式检验,比伦斯(Bierens, 1990)提出对式(8.30)中的 $\mathbf{g}(\mathbf{x})$进行设定的方法,以此确保一致条件矩检验(**consistent conditional moment test**)。但是,如果无法保证检验一致性,就确保它将对特殊的备择假设具有高的势。

8.2.9 *m 检验例子*

为了阐明各种 m 检验,考察 5.2 节引入的泊松回归模型,其泊松密度为 $f(y)=e^{-\mu}\mu^y/y!$,且 $\mu=\exp(\mathbf{x}'\boldsymbol{\beta})$。

对于 $\mathbf{m}(\cdot)$的各种不同选取,我们想要检验:

$$H_0: \mathrm{E}[\mathbf{m}(y,\mathbf{x},\boldsymbol{\beta})]=\mathbf{0}$$

实际上,这个检验将在数据生成过程被设定为泊松密度的假设下进行。

辅助回归

由于通过极大似然法估计,所以能使用 m 检验统计量 M^*,将它计算成 N 倍的源于辅助回归(8.5)非中心化 R^2,其中:

$$1=\hat{\mathbf{m}}(y_i,\mathbf{x}_i,\hat{\boldsymbol{\beta}})'\boldsymbol{\delta}+(y_i-\exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}}))\mathbf{x}_i'\boldsymbol{\gamma}+u_i \tag{8.32}$$

因为 $\hat{\mathbf{s}}=|\partial\ln f(y)/\partial\boldsymbol{\beta}|_{\hat{\beta}}=(y-\exp(\mathbf{x}'\hat{\boldsymbol{\beta}}))\mathbf{x}$,而 $\hat{\boldsymbol{\beta}}$ 是极大似然估计值。在 H_0 下,该检验服从 $\chi^2(\dim(\mathbf{m}))$分布。

一种可供选择的方式是源于辅助回归:

$$1=\hat{\mathbf{m}}(y,\mathbf{x},\mathbf{z},\hat{\boldsymbol{\beta}})'\boldsymbol{\delta}+u \tag{8.33}$$

的统计量 M^{**}。如果 $\mathbf{m}(\cdot)$使得 $\mathrm{E}[\partial\mathbf{m}/\partial\boldsymbol{\beta}]=\mathbf{0}$,那么这个检验渐近地等价于 LM^*,否则它就不服从卡方分布。

矩检验

对条件均值函数的正确设定,即 $\mathrm{E}[y-\exp(\mathbf{x}'\boldsymbol{\beta})|\mathbf{x}]=0$,这可通过:

$$\mathrm{E}[(y-\exp(\mathbf{x}'\boldsymbol{\beta}))\mathbf{z}]=\mathbf{0}$$

的 m 检验来加以确定,其中,**z** 可以是 **x** 的函数。对泊松模型以及其他 LEF 模型来说,**z** 不能等于 **x**,因为 $\hat{\beta}_{ML}$ 的一阶条件利用了约束 $\sum_i(y_i-\exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}}))\mathbf{x}_i'=\mathbf{0}$,从而导致如果 $\mathbf{z}=\mathbf{x}$,那么 M=0。相反,**z** 能包括回归元的平方项与交叉项。

同理,对方差正确设定进行检验,因为泊松分布蕴含条件均值方差等式。由于 $\mathrm{V}[y|\mathbf{x}]-\mathrm{E}[y|\mathbf{x}]=0$ 且 $\mathrm{E}[y|\mathbf{x}]=\exp(\mathbf{x}'\boldsymbol{\beta})$,建议:

$$\mathrm{E}[\{(y-\exp(\mathbf{x}'\boldsymbol{\beta}))^2-\exp(\mathbf{x}'\boldsymbol{\beta})\}\mathbf{x}]=\mathbf{0}$$

的 m 检验。不过,由于 $\mathrm{E}[y|\mathbf{x}]=\exp(\mathbf{x}'\boldsymbol{\beta})$,所以一种变形就是检验:

$$\mathrm{E}[\{(y-\exp(\mathbf{x}'\boldsymbol{\beta}))^2-y\}\mathbf{x}]=\mathbf{0}$$

那么,$\mathbf{m}(\boldsymbol{\beta})=\{(y-\exp(\mathbf{x}\boldsymbol{\beta}))^2-y\}\mathbf{x}$ 具有 $\mathrm{E}[\partial\mathbf{m}/\partial\boldsymbol{\beta}]=\mathbf{0}$ 的性质,因而式(8.7)成立,并且可供选择的回归(8.33)产生了渐近等价于回归(8.23)的检验。

参数模型的标准设定检验是 IM 检验。对于泊松密度,已在式(8.19)定义的 **D** 变成 $\mathbf{D}(y,\mathbf{x},\boldsymbol{\beta})=\{(y-\exp(\mathbf{x}'\boldsymbol{\beta}))^2-y\}\mathbf{x}\mathbf{x}'$,从而我们检验:

$$\mathrm{E}[\{(y-\exp(\mathbf{x}'\boldsymbol{\beta}))^2-y\}\mathrm{Vech}[\mathbf{x}\mathbf{x}']]=\mathbf{0}$$

很明显,就泊松例子而言,IM 检验是对由泊松模型所蕴含的一阶与二阶矩条件进行检验,更一般地讲,对 LEF 模型来说,其结果仍成立。由于此处 $\mathrm{E}[\partial\mathbf{m}/\partial\boldsymbol{\beta}]=\mathbf{0}$,故检验统计量 M^{**} 渐近地等价于 M^{*}。

泊松假设还能利用卡方拟合优度来进行检验。例如,由于在后面模拟例子中的少数几个计数大于三个的胞腔对应于 $y=0,1,2,3$ 或更多一些,在实施检验时,含有 $y=3$ 或更多一些的胞腔将被去掉,因为概率和为 1。因此,对于 $j=0,\cdots,2$ 计算标示变量,当 $y_i=j$ 时,$d_{ij}=1$,否则 $d_{ij}=0$,从而计算出预测概率 $\hat{p}_{ij}=e^{-\hat{\mu}_i}\hat{\mu}_i^j/j!$,其中 $\hat{\mu}_i=\exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}})$。于是,对

$$\mathrm{E}[(\mathbf{d}-\mathbf{p})]=\mathbf{0}$$

进行检验,其中 $\mathbf{d}_i=[d_{i0},d_{i1},d_{i2}]$,而由辅助回归(8.33)知,$\hat{\mathbf{p}}_i=[p_{i0},p_{i1},p_{i2}]$,这里,$\hat{\mathbf{m}}_i=\mathbf{d}_i-\hat{\mathbf{p}}_i$。

模拟结果

数据由泊松模型生成,其均值 $\mathrm{E}[y|\mathbf{x}]=\exp(\beta_1+\beta_2x_2)$,其中,$x_2\sim\mathcal{N}(0,1)$ 且 $(\beta_1,\beta_2)=(0,1)$。对样本量为 200 的样本来说,y 对 **x** 的泊松 ML 回归,得出:

$$\hat{\mathrm{E}}[y|x]=\exp(\underset{(0.089)}{-1.165}+\underset{(0.069)}{1.124}x_2)$$

其中,有关的标准误差已列在小括号中。

各种不同的 m 检验结果已由表 8.1 给出。

表 8.1 泊松回归例子的设定 m 检验[a]

检验类型	H_0其中 $\mu=\exp(\mathbf{x}'\boldsymbol{\beta})$	M*	dof	p 值	M**
1. 正确均值	$\mathrm{E}[(y-\mu)x_2^2]=0$	3.27	1	0.07	0.44
2. 方差=均值	$\mathrm{E}[\{(y-\mu)^2-\mu\}\mathbf{x}]=\mathbf{0}$	2.43	2	0.30	1.89
3. 方差=均值	$\mathrm{E}[\{(y-\mu)^2-y\}\mathbf{x}]=\mathbf{0}$	2.43	2	0.30	2.41
4. 信息矩阵	$\mathrm{E}[\{(y-\mu)^2-y\}\mathrm{Vech}[\mathbf{xx}']]=\mathbf{0}$	2.95	3	0.40	2.73
5. 卡方 GOF	$\mathrm{E}[\mathbf{d}-\mathbf{p}]=\mathbf{0}$	2.50	3	0.48	0.75

[a] y 的数据生成过程是泊松分布，其均值参数为 $\exp(0+x_2)$，样本量 $N=200$。m 检验统计量 M* 服从卡方分布，其自由度已在 dof 列中给出，p 值在 p 值列中给出。一种可供选择的检验统计量 M** 仅对检验 3 和检验 4 是有效的。

考察 IM 检验，举一个利用式(8.32)计算 M* 的例子。由于 $\mathbf{x}=[1,x_2]'$，且 $\mathrm{Vech}[\mathbf{xx}']=[1,x_2,x_2^2]'$，辅助回归是 1 对 $\{(y-\hat{\mu})^2-y\}$、$\{(y-\hat{\mu})^2-y\}x_2$、$\{(y-\hat{\mu})^2-y\}x_2^2$、$(y-\hat{\mu})$、$(y-\hat{\mu})x_2$ 的回归，并得出非中心化 $R^2=0.01473$ 和 $N=200$，这导致 M* = 2.95。相同的 M* 值可直接通过非中心化的解释平方和 2.95 直接获得，同时可间接地作为 N 减去源自这个回归的残差平方和 197.05 而得到。检验统计量服从 $\chi^2(3)$分布，并且 $p=0.40$，因此，零假设在显著性水平 0.05 上没有被拒绝。

对卡方拟合优度检验来说，实际频率分别是 0.435、0.255 和 0.110；而相应的预测频率是 0.429、0.241 和 0.124。这可利用式(8.23)得出，PCGF=0.47，但该统计量已不服从卡方分布，因为它没有控制估计 $\hat{\boldsymbol{\beta}}$中的误差。式(8.27)中的正确统计量卡方拟合优度(CGF)的辅助回归会导致 M* =2.50，它服从卡方分布。

在此模拟研究中，当 M* 的 p 值大于 0.05，全部 5 个矩条件在水平 0.05 上都没有被拒绝。如同人们所料，由于这个模拟例子的数据是由设定密度生成的，所以检验在水平 0.05 上仅有 5%的时间应该被拒绝。一种可供选择的统计量 M** 只有对检验 3 与检验 4 才会有效，那样才有 $\mathrm{E}[\partial\mathbf{m}/\partial\boldsymbol{\beta}]=\mathbf{0}$；否则，它只提供了 M 下界。

8.3 豪斯曼检验

建立在两个不同估计量比较基础上的检验称为豪斯曼检验，以豪斯曼(Hausman, 1978)命名，也称为吴—豪斯曼检验或杜宾—吴—豪斯曼检验。以吴(Wu, 1973)和杜宾(Durbin, 1954)命名，是因为他们都曾提出过类似检验。

8.3.1 *豪斯曼检验*

考察单方程中对回归元内生性的检验。两种可供选择的估计量是 OLS 估计量与 2SLS 估计量，其中，2SLS 估计量为了控制回归元的可能内生性而使用工具。如果存在内生性，那么 OLS 是非一致的，因而这两种估计量将具有不同的概率极限。这就提出了，对内生性进行检验可通过对 OLS 估计量与 2SLS 估计量之差来进行检验，更详细的讨论可参见 8.4.3 节。

更一般地，考察两个估计量 $\hat{\boldsymbol{\theta}}$ 与 $\tilde{\boldsymbol{\theta}}$。下面考察检验：

$$H_0:\ \text{plim}(\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}})=\mathbf{0} \tag{8.34}$$
$$H_a:\ \text{plim}(\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}})\neq\mathbf{0}$$

假定两个根号 N 一致估计量之差在 H_0 下还是根号 N 一致的，其均值为 **0** 且服从极限正态分布，因此：

$$\sqrt{N}(\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}})\xrightarrow{d}\mathcal{N}[\mathbf{0},\mathrm{V_H}]$$

其中，$\mathrm{V_H}$ 表示极限分布中的方差矩阵。于是，豪斯曼检验统计量为：

$$\mathrm{H}=(\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}})'(N^{-1}\hat{\mathrm{V}}_{\mathrm{H}})^{-1}(\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}}) \tag{8.35}$$

在 H_0 下，渐近服从 $\chi^2(q)$ 分布。在水平 α 上，当 $\mathrm{H}>\chi^2_\alpha(q)$ 时，就拒绝 H_0。

在一些应用中，例如对内生性的检验，$\mathrm{V}[\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}}]$ 是小于满秩的形式。于是，把广义逆用于式(8.35)，而且卡方检验具有等于 $\mathrm{V}[\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}}]$ 秩的自由度。

豪斯曼检验能用于参数的子集。例如，关注内容只是可能内生回归元的系数，以及从 OLS 到 2SLS 变动时它是否变化。那么，仅有 $\boldsymbol{\theta}$ 的一个分量被使用，故该检验统计量服从 $\chi^2(1)$ 分布。正如在其他背景下一样，建立在参数子集上的这一检验所得出的结论，不同于建立在全部参数上的检验所得出的结论。

8.3.2 *豪斯曼检验计算*

从原则上讲，计算豪斯曼检验很容易，但在实际应用时得到它很困难，由于需要得到 $\mathrm{V_H}$ 的一致估计值，即 $\sqrt{N}(\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}})$ 的极限方差矩阵。通常有：

$$N^{-1}\mathrm{V_H}=\mathrm{V}[\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}}]=\mathrm{V}[\hat{\boldsymbol{\theta}}]+\mathrm{V}[\tilde{\boldsymbol{\theta}}]-2\mathrm{Cov}[\hat{\boldsymbol{\theta}},\tilde{\boldsymbol{\theta}}] \tag{8.36}$$

前两个量均容易由通常输出结果计算，但第三个量则不能。

在零假设下计算完全有效估计量

尽管豪斯曼检验的基本零假设与备择假设如同式(8.34)一样，但在应用时要记住，通常存在特定的零假设模型及备择假设。例如，在比较 OLS 估计量与 2SLS 估计量时，零假设模型则允许一些回归元为内生的。

若 $\hat{\boldsymbol{\theta}}$ 是零假设模型的有效估计量，则 $\mathrm{Cov}[\hat{\boldsymbol{\theta}},\tilde{\boldsymbol{\theta}}]=\mathrm{V}[\hat{\boldsymbol{\theta}}]$。其证明参见习题 8.3。这蕴含 $\mathrm{V}[\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}}]=\mathrm{V}[\tilde{\boldsymbol{\theta}}]-\mathrm{V}[\hat{\boldsymbol{\theta}}]$，因此：

$$\mathrm{H}=(\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}})'(\hat{\mathrm{V}}[\tilde{\boldsymbol{\theta}}]-\hat{\mathrm{V}}[\hat{\boldsymbol{\theta}}])^{-1}(\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}}) \tag{8.37}$$

该统计量因仅需要参数 $\hat{\boldsymbol{\theta}}$ 与 $\tilde{\boldsymbol{\theta}}$ 估计渐近方差矩阵，所以具有相当多的优点。使用允许保留参数、方差矩阵估计值并利用矩阵命令的计算程序，这样做是有益的。

例如，如果假设误差是同方差的，那么这种简化能应用到线性回归模型应用中的内生性检验上。于是，$\hat{\boldsymbol{\theta}}$ 在没有内生性的零假设下成为完全有效的 OLS 估计量，而 $\tilde{\boldsymbol{\theta}}$ 是 2SLS 估计量。可是，需要小心谨慎，以便确保方差矩阵的一致估计值使得 $\mathrm{V}[\tilde{\boldsymbol{\theta}}]-\mathrm{V}[\hat{\boldsymbol{\theta}}]$ 是正定的[参见鲁德(Rudd, 1984)]。在 OLS 与 2SLS 的比较中，方差矩阵估计量 $\hat{\mathrm{V}}[\tilde{\boldsymbol{\theta}}]$ 及 $\hat{\mathrm{V}}[\hat{\boldsymbol{\theta}}]$ 应使用误差方差 σ^2 的相同估计值。

尤其是，当 θ 是纯量的或对参数向量的唯一一个分量进行检验时，豪斯曼检验(8.37)的形式可通过手工方式很容易地计算。于是：

$$\mathrm{H}=(\hat{\theta}-\tilde{\theta})^2/(\tilde{s}^2-\hat{s}^2)$$

服从 $\chi^2(1)$分布，其中，$\hat{s}$ 与 $\tilde{s}$ 都表示 $\hat{\theta}$ 与 $\tilde{\theta}$ 的报告标准误差。

辅助回归

在一些重要情况下，豪斯曼检验更简单地计算成为在增广 OLS 回归中对回归元子集的显著性的标准检验，推导是在 $\hat{\boldsymbol{\theta}}$ 为完全有效的假设下进行的。

一些例子将在 8.4.3 节和 21.4.3 节给出。

稳健豪斯曼检验

豪斯曼检验的较简单形式(8.37)以及标准的辅助回归，都需要 $\hat{\boldsymbol{\theta}}$ 是完全有效的强分布假设。与在相对弱分布假设下实施的稳健推断方法相比，这是其对立情况。

从原则上讲，对 $\mathrm{Cov}[\hat{\boldsymbol{\theta}},\tilde{\boldsymbol{\theta}}]$可直接估计，从而可估计出 $\mathrm{V_H}$。假定 $\hat{\boldsymbol{\theta}}$ 与 $\tilde{\boldsymbol{\theta}}$ 是求解$\sum_i \mathbf{h}_{1i}(\hat{\boldsymbol{\theta}})=\mathbf{0}$ 与$\sum_i \mathbf{h}_{2i}(\tilde{\boldsymbol{\theta}})=\mathbf{0}$ 的 m 估计量。定义 $\hat{\boldsymbol{\delta}}'=[\hat{\boldsymbol{\theta}},\tilde{\boldsymbol{\theta}}]$。于是，$\mathrm{V}[\hat{\boldsymbol{\delta}}]=\mathbf{G}_0^{-1}\mathbf{S}_0(\mathbf{G}_0^{-1})'$，其中，$\mathbf{G}_0$ 与 $\mathbf{S}_0$ 已由 6.6 节定义，其简化形式为 $\mathbf{G}_{12}=\mathbf{0}$。人们期望，$\mathrm{V}[\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}}]=\mathrm{RV}[\hat{\boldsymbol{\delta}}]\mathbf{R}'$，其中，$\mathbf{R}=[\mathbf{I}_q,-\mathbf{I}_q]$。实施起来需要应用于特定情况的额外编程。

一种较简单的方法是自助法(参见 11.6.3 节)，尽管在一些应用中需要小心谨慎，以便确保在卡方检验时有正确的自由度。

另一种非完全有效 $\hat{\boldsymbol{\theta}}$ 的方法是使用辅助回归，该辅助回归在有效情况下是适宜的，但为了实施回归元子集检验而利用稳健标准误差。这种稳健检验可直接实施，并在对关注的错误设定进行检验时具有势，尽管它可能不一定是等价于使用由式(8.35)给出 H 的更一般形式的豪斯曼检验。21.4.3 节将给出一个例子。

最后，计算出一些界，这并不需要计算 $\mathrm{Cov}[\hat{\boldsymbol{\theta}},\tilde{\boldsymbol{\theta}}]$。对纯量随机变量来说，$\mathrm{Cov}[x,y]\leqslant s_x s_y$。就纯量情况而言，这提出 H 的上界$(\hat{\theta}-\tilde{\theta})^2/(\hat{s}^2+\tilde{s}^2-2\hat{s}\tilde{s})$，其中，$\hat{s}^2=\hat{\mathrm{V}}[\hat{\theta}]$与 $\tilde{s}^2=\hat{\mathrm{V}}[\tilde{\theta}]$。在 $\hat{\theta}$ 与 $\tilde{\theta}$ 是正相关的假设下，H 的下界是 $N(\hat{\theta}-\tilde{\theta})/(\hat{s}^2+\tilde{s}^2)$。不过，实际应用中，这些界是相当广泛的。

8.3.3 *豪斯曼检验的势*

豪斯曼检验是非常一般的方法，该方法没有显性地表述一种可供选择的假设，因此，不需要对特殊可供选择假设具有高的势。

例如，考察完全参数模型对排除性约束的检验。注意到，零假设 H_0：$\boldsymbol{\theta}_2=\mathbf{0}$，其中，$\boldsymbol{\theta}$ 被分割成$(\boldsymbol{\theta}_1',\boldsymbol{\theta}_2')'$。一个明显设定检验是 $\hat{\boldsymbol{\theta}}_1-\tilde{\boldsymbol{\theta}}_1$之差的豪斯曼检验，其中，$(\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_2)$表示无约束 MLE，而$(\tilde{\boldsymbol{\theta}}_1,\mathbf{0})$表示 $\boldsymbol{\theta}$ 的约束 MLE。霍利(Holly, 1982)已经证明，这个豪斯曼检验与 H_0：$\mathcal{I}_{11}^{-1}\mathcal{I}_{12}\boldsymbol{\theta}_2=\mathbf{0}$ 的经典检验(沃尔德、LR 或 LM)是一样的，其中，$\mathcal{I}_{ij}=\mathrm{E}[\partial^2\mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)/\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j]$，而不是 H_0：$\boldsymbol{\theta}_2=\mathbf{0}$ 的情况。如果 $\mathcal{I}_{12}$是列满秩的且 $\dim(\boldsymbol{\theta}_1)\geqslant\dim(\boldsymbol{\theta}_2)$，那么这两种检验是一样的，进而 $\mathcal{I}_{11}^{-1}\mathcal{I}_{12}\boldsymbol{\theta}_2=\mathbf{0}$ 当且仅当 $\boldsymbol{\theta}_2=\mathbf{0}$。否则，它们是不同的。很明显，当信息矩阵是分块对角时，豪斯曼检验将

不具有对 H_0 的势，进而 $\mathcal{I}_{12}=\mathbf{0}$。霍利（Holly，1987）将这种分析推广到非线性假设。

8.4 对某些普遍错误设定的检验

本节阐述对某些普遍模型错误设定的检验。关注内容在于能利用辅助回归进行计算的检验统计量，这就可利用最少的假设实施对异方差误差稳健的推断。

8.4.1 *对省略变量检验*

除特殊情况之外，省略变量通常会导致非一致的参数估计，例如，线性模型中省略变量不与其他回归元相关。因此，重要的是检验潜在省略变量。

最经常使用的是沃尔德检验，与估计含有排除省略变量的约束模型相比，估计含有省略变量的模型通常不再困难。进一步地，这个检验可使用稳健三明治标准误差，只有当稳健三明治误差是必需的时，估计量仍是一致性的，这样做才真正才有意义。

把关注限制在 ML 估计，一种可供选择的方法是，估计具有潜在不相关的回归元模型与没有潜在不相关回归元的模型，然后实施 LR 检验。

在某些背景下，很容易计算 LM 检验的稳健形式。例如，考察均值为 $\exp(\mathbf{x}_1'\boldsymbol{\beta}_1+\mathbf{x}_2'\boldsymbol{\beta}_2)$ 的泊松模型的 H_0：$\boldsymbol{\beta}_2=\mathbf{0}$ 的检验。此 LM 检验统计量是建立在得分统计量 $\sum_i \mathbf{x}_i\tilde{u}_i$ 的基础上的，其中，$\tilde{u}_i=y_i-\exp(\mathbf{x}_{1i}'\tilde{\boldsymbol{\beta}}_1)$（参见 7.3.2 节）。现在，关于 $N^{-1/2}\sum_i \mathbf{x}_i u_i$ 方差的异方差性稳健估计值是 $N^{-1}\sum_i u_i^2\mathbf{x}_i\mathbf{x}_i'$，其中，$u_i=y_i-\mathrm{E}[y_i|\mathbf{x}_i]$，可以证明：

$$\mathrm{LM}^{+}=\Big[\sum_{i=1}^{n}\mathbf{x}_i\,\tilde{u}_i\Big]'\Big[\sum_{i=1}^{n}\tilde{u}_i^2\mathbf{x}_i\mathbf{x}_i'\Big]^{-1}\Big[\sum_{i=1}^{n}\mathbf{x}_i\,\tilde{u}_i\Big]$$

是稳健 LM 检验统计量，该统计量不需要在 H_0 下的 $\mathrm{V}[u_i|\mathbf{x}_i]=\exp(\mathbf{x}_{1i}'\boldsymbol{\beta}_1)$ 泊松约束。这能计算为源于 1 对 $\mathbf{x}_{1i}\tilde{u}_i$ 及 $\mathbf{x}_{2i}\tilde{u}_i$ 回归的未中心化R^2 的 N 倍。更一般地讲，对线性指数族中假定的模型来说，这类稳健 LM 检验是可行的，因为这种模型中的得分统计量再次是残差 $\tilde{u}_i$ 的加权平均[参见伍德里奇（Wooldridge，1991）]。这一类包括 OLS，而当通过 2SLS 或 NLS 进行估计时，可适当修改，参见伍德里奇（Wooldridge，2002）。

8.4.2 *异方差性检验*

在存在异方差性时，由最小二乘或工具变量方法估计的条件均值线性或非线性回归模型中的参数估计值，保持它们的一致性。唯一需要校正的是这些估计值的标准误差。这并不需要对异方差性进行建模，因为在最少分布的假设下，异方差稳健标准误差可利用怀特（White，1982）的结果加以计算。因此，对异方差性很少需要进行检验，除非估计量的有效性是重点关注的内容。不过，我们对异方差性检验的一些结果加以归纳总结。

我们以线性回归模型 $y=\mathbf{x}'\boldsymbol{\beta}+u$ 的 LS 估计开始。假定异方差性可由 $\mathrm{V}[u|\mathbf{x}]=g(\alpha_1+\mathbf{z}'\boldsymbol{\alpha}_2)$进行建模，其中，$\mathbf{z}$ 通常表示 $\mathbf{x}$ 的子集，而 $g(\cdot)$常常表示指数函数。文献关注于利用 LM 方法对 H_0：$\boldsymbol{\alpha}_2=\mathbf{0}$ 进行的检验，因为与沃尔德及 LR 检验不同，这仅仅要求 $\boldsymbol{\beta}$的 OLS 估计。布鲁什和帕甘(Breusch and Pagan，1979)的标准 LM 检验紧密依赖于正态分布误差的假设，因为它使用在 H_0 下的 $\mathrm{E}[u^4|\mathbf{x}^4]=3\sigma^4$ 的约束。凯恩克(Koenker，1981)曾经提出 LM 检验的更稳健形式，源于$\hat{u}_i$ 对 1 与 $\mathbf{z}_i$ 回归的 NR^2，其中，$\hat{u}_i$ 表示 OLS 残差。该检验需要较弱的假设——$\mathrm{E}[u^4|\mathbf{x}]$是常数。像布鲁什—帕甘检验一样，它对函数 $g(\cdot)$的选择而言是不变的。异方差性的怀特(White，1980a)检验等价于这个 LM 检验，满足 $\mathbf{z}=\mathrm{Vech}[\mathbf{x}\mathbf{x}']$。该检验能被进一步推广到令 $\mathrm{E}[u^4|\mathbf{x}]$随 $\mathbf{x}$ 而变化的情况，尽管常值对检验的假设而言是有道理的，因为 H_0 已经设定，$\mathrm{E}[u^2|\mathbf{x}]$是一个常值。

对条件均值的非线性模型来说，在性质上可完成类似结果，这里的非线性模型假定对于错误设定异方差性的一种特殊形式进行检验。例如，泊松回归模型设 $\mathrm{V}[y|\mathbf{x}]=\exp(\mathbf{x}'\boldsymbol{\beta})$。更一般地讲，对线性指数族模型而言，尽管错误设定的异方差性及性质类似于此处应用的结果，但准 MLE 是一致的。于是，倘若可以使用 5.7.4 节阐述的稳健标准误差，即使关于异方差性的模型被错误设定，但获得有效推断是可能的。如果人们还希望对异方差性的正确设定进行检验，那么稳健 LM 检验是可行的[参见伍德里奇(Wooldridge，1991)]。

在一些非线性模型中，异方差性能导致参数估计值非一致性更为严重的后果。一个重要例子是 Tobit 模型(参见第 16 章)，含有正态同方差误差的线性回归模型由于删失或截取而变成非线性的，于是，对异方差性进行检验变得更加重要。对 $\mathrm{V}[u|\mathbf{x}]$模型可加以设定，也可实施沃尔德检验、LR 检验或 LM 检验，或使用关于异方差性的 m 检验[参见帕甘和维拉(Pagan and Vella，1989)]。

8.4.3 *内生性豪斯曼检验*

工具变量估计量应该仅在需要它们时才好使用，因为倘若所有回归元都是外生的，则最小二乘法估计量就是更有效的，并且由 4.9 节知，这种有效性的损失是相当大的。因此，检验是否需要工具变量方法是有用的。对回归元内生性的检验(**test for endogeneity of regressor**)是，将工具变量估计与最小二乘法估计进行对比。若回归元是内生的，则在极限形式上，这些估计值将会有所不同；而若回归元是外生的，两种估计量将会一样。因此，最小二乘法与工具变量估计值之间的差异能解释成内生性的证据。

这个例子提供了豪斯曼检验的最初动机。考察线性回归模型：

$$y=\mathbf{x}_1'\boldsymbol{\beta}_1+\mathbf{x}_2'\boldsymbol{\beta}_2+u \tag{8.38}$$

其中，$\mathbf{x}_1$ 表示潜在内生的，$\mathbf{x}_2$ 表示外生的。设 $\hat{\boldsymbol{\beta}}$ 表示式(8.38)中的 OLS 估计量，而$\tilde{\boldsymbol{\beta}}$ 表示式(8.38)中的 2SLS 估计量。一旦假定同方差误差，则 OLS 在没有内生性的零假设下是有效的，内生性的豪斯曼检验，可利用式(8.37)中定义的检验统计量 H 加以计算。因为可以证明，$\mathrm{V}[\hat{\boldsymbol{\beta}}]-V[\tilde{\boldsymbol{\beta}}]$不是满秩的，但需要广义逆，并且

自由度是 $\dim(\boldsymbol{\beta}_1)$ 而不是 $\dim(\boldsymbol{\beta})$。

豪斯曼(Hausman，1978)证明，在增广 OLS 回归：

$$y=\mathbf{x}_1'\boldsymbol{\beta}_1+\mathbf{x}_2'\boldsymbol{\beta}_2+\hat{\mathbf{x}}_1'\boldsymbol{\gamma}+u$$

中，通过对 $\boldsymbol{\gamma}=\mathbf{0}$ 的检验来更简单地实施检验，其中，$\hat{\mathbf{x}}_1$ 表示源自 $\mathbf{x}_1$ 对工具 $\mathbf{z}$ 的多元回归简化式中内生回归元 $\mathbf{x}_1$ 的预测值。等价地，我们能在以下增广 OLS 回归中检验 $\boldsymbol{\gamma}=\mathbf{0}$：

$$y=\mathbf{x}_1'\boldsymbol{\beta}_1+\mathbf{x}_2'\boldsymbol{\beta}_2+\hat{\mathbf{v}}_1'\boldsymbol{\gamma}+u$$

其中，$\hat{\mathbf{v}}_1$ 表示源自 $\mathbf{x}_1$ 对工具 $\mathbf{z}$ 的多元回归简化式的残差。就这些检验而言，从直观上看，如果式(8.38)中的 u 与 $\mathbf{x}_1$ 及 $\mathbf{x}_2$ 不相关，那么 $\boldsymbol{\gamma}=\mathbf{0}$。相反，如果 u 与 $\mathbf{x}_1$ 相关，那么这将由 $\mathbf{x}_1$ 的其他变换譬如 $\hat{\mathbf{x}}_1$ 及 $\hat{\mathbf{v}}_1$ 的显著性进行处理。

对横截面数据来说，一种习惯做法是假定异方差误差。那么，式(8.38)的 OLS 估计量 $\hat{\boldsymbol{\beta}}$ 是无效的，而且不能使用豪斯曼检验的较简单形式(8.37)。不过，倘若利用方差矩阵的异方差一致估计来对 $\boldsymbol{\gamma}=\mathbf{0}$ 加以检验，则前面的增广 OLS 回归还是能使用的。实际上，这应该等价于豪斯曼检验，因为由戴维森和麦金农(Davidson and MacKinnon，1993，第 239 页)的讨论知道，这些增广回归的 $\hat{\boldsymbol{\gamma}}_{\text{OLS}}$ 等于 $\mathbf{A}_N(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})$，其中，$\mathbf{A}_N$ 表示满秩矩阵且具有有限概率极限。

可能有另外的内生性豪斯曼检验。假定 $y=\mathbf{x}_1'\boldsymbol{\beta}_1+\mathbf{x}_2'\boldsymbol{\beta}_2+\mathbf{x}_3'\boldsymbol{\beta}_3+u$，其中，$\mathbf{x}_1$ 表示潜在内生的，假定 $\mathbf{x}_2$ 是内生的，并假定 $\mathbf{x}_3$ 是外生的。于是，$\mathbf{x}_1$ 的内生性能通过把仅含有 $\mathbf{x}_2$ 工具的 2SLS 估计量与既含有 $\mathbf{x}_1$ 又含有 $\mathbf{x}_2$ 工具的 2SLS 估计量加以比较。还可将豪斯曼检验推广到非线性回归模型上，只是要用 NLS 代替 OLS，并用 NL2SLS 代替 2SLS。戴维森和麦金农(Davidson and MacKinnon，1993)曾阐述，一旦假定同方差误差，则增广回归能用于计算有关豪斯曼检验的情况。当 $\hat{\boldsymbol{\theta}}$ 不是有效估计量时，包括计算 $\mathrm{V}[\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}}]$ 例子，姆罗茨(Mroz，1987)已经提供一个好的内生性检验应用。

8.4.4 外生性的 OIR 检验

如果使用工具变量估计量，那么为使工具变量估计量成为一致的，所用工具必须是外生的。对恰好识别模型来说，检验工具外生性是不可能的。不过，需要使用先验理由来判断工具有效性。4.8.2 节已给出一些例子。但是，就过度识别模型而言，对工具外生性进行检验是可能的。

我们以线性回归开始。于是，$y=\mathbf{x}'\boldsymbol{\beta}+u$，若 $\mathrm{E}[u|\mathbf{z}]=\mathbf{0}$ 或 $\mathrm{E}[\mathbf{z}u]=\mathbf{0}$，则工具 $\mathbf{z}$ 是有效的。H_0：$\mathrm{E}[\mathbf{z}u]=\mathbf{0}$ 的一个明显检验是建立在 $N^{-1}\sum_i\mathbf{z}_i\hat{u}_i$ 背离 0 的基础上。在恰好识别的情况下，工具变量估计量是 $N^{-1}\sum_i\mathbf{z}_i\hat{u}_i=0$ 的解，所以这个检验没有用。在过度识别的情况下，6.3.8 节曾阐述的过度识别约束检验是：

$$\mathrm{OIR}=\hat{\mathbf{u}}'\mathbf{Z}\hat{\mathbf{S}}^{-1}\mathbf{Z}'\hat{\mathbf{u}} \tag{8.39}$$

其中，$\hat{\mathbf{u}}=\mathbf{y}-\mathbf{x}\hat{\boldsymbol{\beta}}$，$\hat{\boldsymbol{\beta}}$ 表示对 $\mathbf{u}'\mathbf{Z}\hat{\mathbf{S}}^{-1}\mathbf{Z}'\mathbf{u}$ 求极小值的最优广义矩方法估计量，而 $\hat{\mathbf{S}}$ 表示关于 $\mathrm{plim}\,N^{-1}\sum_i u_i^2\mathbf{z}_i\mathbf{z}_i'$ 是一致的。汉森(Hansen，1982)的 OIR 检验是将萨根

(Sargan, 1958)提出的检验推广到线性工具变量上,而且检验统计量(8.39)经常称为萨根检验(**Sargan test**)。当 OIR 很大时,就拒绝矩条件,从而工具变量估计量是非一致的。对 H_0 拒绝,通常可以解释为工具 $\mathbf{z}$ 是内生的证据,但它也可以是模型错误设定的证据,因此,实际上 $y\neq\mathbf{x}'\boldsymbol{\beta}+u$。在上述任何一种情况下,拒绝均表明,工具变量估计量是有问题的。

正如 6.3.9 节正式推导的,OIR 在 H_0 下服从 $\chi^2(r-K)$分布,其中,$(r-K)$表示过度识别约束的个数。为了获得此结果的某种直观理解,专门研究同方差误差是有用的。于是,$\hat{\mathbf{S}}=\hat{\sigma}^2\mathbf{Z}'\mathbf{Z}$,其中,$\hat{\sigma}^2=\hat{\mathbf{u}}'\hat{\mathbf{u}}/(N-K)$,所以:

$$\mathrm{OIR}=\frac{\hat{\mathbf{u}}'\mathbf{P}_{\mathbf{Z}}\hat{\mathbf{u}}}{\hat{\mathbf{u}}'\hat{\mathbf{u}}/(N-K)}$$

其中,$\mathbf{P}_{\mathbf{Z}}=\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$。因此,OIR 是关于 $\hat{\mathbf{u}}$ 的二次形式之比。在 H_0 下,分子具有概率极限 $\sigma^2(r-K)$,而分母具有 $\mathrm{plim}\,\hat{\sigma}^2=\sigma^2$。因此,该比值是以 $r-K$ 为中心的,但这是 $\chi^2(r-K)$随机变量的均值。

式(8.39)中的检验统计量,如同 6.5 节一样,通过直接定义 $u_i=y-g(\mathbf{x},\boldsymbol{\beta})$或 $u=r(y,\mathbf{x},\boldsymbol{\beta})$,可立刻扩展到非线性回归上,而对于线性方程组与面板估计量,则要对 $\mathbf{u}$ 适当定义(参见 6.9 节和 6.10 节)。

对具有同方差误差的线性工具变量来说,可提出一种对式(8.39)可供选择的 OIR 检验。玛格达利诺斯(Magdalinos, 1988)曾经对这些检验加以比较。人们还能使用过度约束子集的增量 OIR 检验。

8.4.5 *RESET* 检验

一种普遍的函数形式错误设定,可能涉及被忽略的某些回归元的非线性。考察回归 $\mathbf{y}=\mathbf{x}'\boldsymbol{\beta}+\mathbf{u}$,其中,我们假定回归元以线性方式进入,且与误差 $\mathbf{u}$ 是渐近不相关的。为了检验非线性,一种简单的方法是引入外生变量的幂函数,比如最普遍的是,平方项作为额外的独立回归元,同时利用沃尔德检验或 F 检验,对这些额外变量的统计显著性加以检验。这要求研究者具有特定的理由考虑非线性,很明显,该种方法对分类变量不起作用。

拉姆齐(Ramsey, 1969)提出,对回归省略变量的检验,能系统地表示为对函数形式的检验。此建议是针对最初回归加以拟合,并生成作为拟合值 $\hat{\mathbf{y}}=\mathbf{x}'\hat{\boldsymbol{\beta}}$ 的新回归元,比如 $\mathbf{w}=[(\mathbf{x}'\hat{\boldsymbol{\beta}})^2,\ (\mathbf{x}'\hat{\boldsymbol{\beta}})^3,\cdots,\ (\mathbf{x}'\hat{\boldsymbol{\beta}})^p]$。然后,估计模型 $\mathbf{y}=\mathbf{x}'\boldsymbol{\beta}+\mathbf{w}'\boldsymbol{\gamma}+\mathbf{u}$,并且对非线性的检验是 p 个约束的沃尔德检验,H_0: $\boldsymbol{\gamma}=\mathbf{0}$ 对 H_0: $\boldsymbol{\gamma}\neq\mathbf{0}$。具体地讲,使用小 p 值,诸如 2 或 3。这个检验对异方差性来说是稳健的。

8.5 区分嵌套模型

当一个模型是另一个模型的特殊情况,则称这两个模型是嵌套的(**nested**);当两个模型中的任何一个都不能表述成另一个的特殊情况,则称两个模型是非嵌套的(**nonnested**)。利用参数约束的标准假设检验,即把一个模型简化成另外一个,对嵌套模型进行区分是可能的。不过,在非嵌套模型的情况下,需要发展一些可供选

择的方法。

这部分内容关注似然框架下对非嵌套模型的辨别，有关结果已得到很好的研究。8.5.4节将给出非似然情况的简要讨论。辨别模型的贝叶斯方法，将在13.8节加以阐述。

8.5.1 *信息准则*

信息准则是含有自由度调整的对数似然准则。具有最小信息准则的模型是人们所偏爱的。

一种基本直观理解如下，当用极大化对数似然值加以测量，并以支持简单模型的简约性原理作为尺度，则模型拟合间就存在矛盾。模型拟合能通过增加模型复杂性来得以改进。不过，如果所得到的拟合改进能充分补偿简约性的损失，那么只须添加参数。注意到，依照此观点看，正在研究的模型集合应该包括“真实数据生成过程”就没有必要了。各种不同的信息准则会随着准则处罚模型的复杂性程度不同而变化。

赤池(Akaike, 1973)最初提出赤池信息准则(**Akaike information criterion**)：

$$\mathrm{AIC}=-2\ln L+2q \tag{8.40}$$

其中，q 表示参数的个数，具有最小 AIC 的模型是人们所偏爱的。运用信息准则这一术语，是因为雨宫(Amemiya, 1980)更简单阐述的基础理论可利用库尔贝克—利布勒(Kullback-Liebler information criteria，简记为 KLIC)信息准则对不同模型加以区别。

人们提出了对 AIC 的相当多的改进，所有的 $-2\ln L+g(q,N)$ 形式都是关于设定罚函数 $g(\cdot)$ 大于 $2q$ 的。一种最流行的变形是贝叶斯信息准则(**Bayesian information criteria**)：

$$\mathrm{BIC}=-2\ln L+(\ln N)q \tag{8.41}$$

它是由施瓦茨(Schwarz, 1978)提出的。施瓦茨假定 y 具有参数为 $\boldsymbol{\theta}$ 的指数族密度，第 j 个模型具有参数 $\boldsymbol{\theta}_j$，满足 $\dim[\boldsymbol{\theta}_j]=q_j<\dim[\boldsymbol{\theta}]$，而且先验的不同模型是关于每个 $\boldsymbol{\theta}_j$ 先验的加权和。施瓦茨已经证明，在这些假设下，对后验概率求极大值(参见第13章)，渐近地等价于选取模型，使其 $\ln L-(\ln N)q_j/2$ 最大化。由于这等价于求式(8.41)的极小值，所以施瓦茨方法称为贝叶斯信息准则。建立在类似于 BIC 的对 KLIC 求极小值基础上的 AIC 精炼是一致 AIC(**consistent AIC**)，即 $\mathrm{CAIC}=-2\ln L+(1+\ln N)q$。一些作者通过对式(8.40)及式(8.41)的右边除以 N，来定义譬如 AIC 与 BIC 等的准则。

假如模型简约性重要，BIC 能更广泛地用作模型水平(**model-size**)惩罚，这是因为 AIC 相对更小。考察分别具有参数 q_1 与 q_2 的两个嵌套模型，其中，$q_2=q_1+h$。那么，实施 LR 检验是可能的，并且当 $2\ln L$ 增加到 $\chi^2_{0.05}(h)$ 时，在显著性水平5%上支持较大的模型。当 $2\ln L$ 增加多于 $2h$ 时，AIC 支持较大的模型，当 $h<7$ 时，与 LR 检验相比，其模型水平惩罚较少。特别地，对于 $h=1$，也就是一个约束，LR 检验使用5%临界值3.84，而 AIC 则使用更小值2。当 $2\ln L$ 增加到 $h\ln N$ 时，BIC

支持较大的模型，与 AIC 或水平为 0.05 的 LR 检验相比，其惩罚更大一些（除非 N 格外小）。

贝叶斯信息准则会随着样本量增加而增大惩罚，而传统假设检验在诸如 5% 的显著性水平上则不会这样。对具有 $q_2=q_1+1$ 的嵌套模型来说，以较小的 BIC 为基础选取较大模型，等价于对 $N=10^2$、10^4 以及 10^6 利用双侧 t 检验的临界值 $\sqrt{\ln N}$进行检验，此时它们分别等于 2.15、3.03 以及 3.72。通过对比，具有水平 0.05 的传统假设检验使用了未变化的临界值 1.96。更一般地讲，对服从 $\chi^2(h)$分布检验统计量而言，BIC 建议，利用 $h\ln N$ 的临界值而不是通常的 $\chi^2_{0.05}(h)$。

给定模型简单性，惩罚似然准则经常用于选取“最佳模型”。不过，即使存在，至少哪种准则应该受到人们的偏爱，这一点并没有清晰答案。在推导 AIC 以及有关测算时，涉及相当程度的近似，并且损失函数而不是对 KLIC 求极小值，或者在 BIC 情况下对后验概率求极大值或许更合适。从决策理论的观点来看，从模型集合中选取模型，应该依赖于模型的使用意图，例如，模型目的是归纳复杂现实性的主要特性，或者预测某些结果，或检验某个重要的假设。在应用研究中，很难看出对经济计量模型使用意图的明确阐述。

8.5.2 *非嵌套模型的 Cox 似然比检验*

考察在两个参数模型之间进行选取的问题。设模型 F_θ 具有密度 $f(y|\mathbf{x},\boldsymbol{\theta})$，模型 G_γ 具有密度 $g(y|\mathbf{x},\boldsymbol{\gamma})$。

模型 F_θ 对模型 G_γ 的似然比检验，建立在下式基础上：

$$\mathrm{LR}(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\gamma}})\equiv\mathcal{L}_f(\hat{\boldsymbol{\theta}})-\mathcal{L}_g(\hat{\boldsymbol{\gamma}})=\sum_{i=1}^{N}\ln\frac{f(y_i|\mathbf{x}_i,\hat{\boldsymbol{\theta}})}{g(y_i|\mathbf{x}_i,\hat{\boldsymbol{\gamma}})}\tag{8.42}$$

当 G_γ 嵌套在 F_θ 之中，则由 7.3.1 节知，$2\mathrm{LR}(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\gamma}})$在零假设 $F_\theta=G_\gamma$ 下服从卡方分布。然而，当模型是非嵌套时，该结果不再成立。

考克斯(Cox，1961，1962b)在 F_θ 是真实模型但模型是非嵌套的特殊情况下，通过在 F_θ 是真实模型的假设下应用中心极限定理，求解了这个问题。

若不能在解析形式上得出 $\mathrm{E}_f[\ln(f(y|\mathbf{x},\boldsymbol{\theta})/g(y|\mathbf{x},\boldsymbol{\gamma}))]$，其中，$\mathrm{E}_f$ 表示关于密度 $f(y|\mathbf{x},\boldsymbol{\theta})$的期望，则这一方法在计算上就很难实施。进一步地，如果类似的检验统计量可借助于对 F_θ 与 G_γ 的作用相反来获得，那么既可能求出模型 F_θ 被拒绝而支持 G_γ，又可能求出模型 G_γ 被拒绝而支持 F_θ。因此，检验不一定是对模型选择的检验，因为它不一定选取一个或另一个；反之，具体说，会出现没有一个模型通过设定检验、一个模型通过设定检验，或者两个模型都通过设定检验的情况。

在一些情况下，可获得考克斯统计量的解析形式。非嵌套线性回归模型 $y=\mathbf{x}'\boldsymbol{\beta}+u$ 与 $y=\mathbf{z}'\boldsymbol{\gamma}+v$ 具有同方差正态分布误差[参见佩萨兰(Pesaran，1974)]。对于非嵌套变换模型 $h(y)=\mathbf{x}'\boldsymbol{\beta}+u$ 与 $g(y)=\mathbf{z}'\boldsymbol{\gamma}+v$，其中，$h(y)$与$g(y)$都是已知变换，参见佩萨兰和佩萨兰(Pesaran and Pesaran，1995)，他们均使用基于模拟方法。例如，这允许对线性参数模型与对数线性参数模型加以区分，这里，$h(\cdot)$为恒等变换，而 $g(\cdot)$为对数变换。佩萨兰和佩萨兰(Pesaran and Pesaran，1995)将

该思想用到第 14 章阐述的对 logit 模型与 probit 模型的选取上。

8.5.3 非嵌套模型翁似然比检验

翁(Vuong, 1989)曾提出 LR 检验统计量的非常一般的分布理论,它既涵盖嵌套模型又涵盖非嵌套模型,并且更为显著的是,允许数据生成过程成为既不同于 $f(\cdot)$ 又不同于 $g(\cdot)$ 的未知密度。

此处阐述翁的渐近结果,有助于理解翁在其论文中阐明的各种检验,这些渐近结果相对很复杂,如同在一些情况下,检验统计量是含有权数的卡方和,其权数很难计算出来。

翁提出以下检验:

$$H_0: \mathrm{E}_0\left[\ln \frac{f(y|\mathbf{x},\boldsymbol{\theta})}{g(y|\mathbf{x},\boldsymbol{\gamma})}\right]=0 \tag{8.43}$$

其中,E_0 表示关于真实数据生成过程 $h(y|\mathbf{x})$ 的期望,此真实数据生成过程可能是未知的。这等价于去检验 $\mathrm{E}_h[\ln(h/g)]-E_h[\ln(h/f)]=0$,或检验 f 与 g 两个密度是否具有相同的库尔贝克—利布勒信息准则(参见 5.7.2 节)。就 H_f: $\mathrm{E}_0[\ln(f/g)]>0$ 与 H_g: $\mathrm{E}_0[\ln(f/g)]<0$ 而言,可能具有一种单侧的可供选择方案。

H_0 的一个明显检验是,式(8.42)定义的样本类似形式 $\mathrm{LR}(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\gamma}})$ 是否异于 0 的 m 检验。此处,检验统计量的分布可借助于可能未知的数据生成过程来获得。这样做是可行的,因为由 5.7.1 节,准 MLE $\hat{\boldsymbol{\theta}}$ 收敛到伪真值 $\boldsymbol{\theta}^*$,并且 $\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)$ 服从极限正态分布,准 MLE $\hat{\boldsymbol{\gamma}}$ 具有类似结果。

一般性结果

所得到的 $\mathrm{LR}(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\gamma}})$ 分布,依据两个模型在 $f(y|\mathbf{x},\boldsymbol{\theta}_*)=g(y|\mathbf{x},\boldsymbol{\gamma}_*)$ 的意义上是否等价而变化,其中,$\boldsymbol{\theta}_*$ 与 $\boldsymbol{\gamma}_*$ 分别表示 $\boldsymbol{\theta}$ 与 $\boldsymbol{\gamma}$ 的伪真实值,这两个模型可能都不正确。

当 $f(y|\mathbf{x},\boldsymbol{\theta}_*)=g(y|\mathbf{x},\boldsymbol{\gamma}_*)$ 时,有:

$$2\mathrm{LR}(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\gamma}}) \xrightarrow{d} M_{p+q}(\boldsymbol{\lambda}_*) \tag{8.44}$$

其中,p 与 q 分别表示 $\boldsymbol{\theta}$ 与 $\boldsymbol{\gamma}$ 的维数,而 $M_{p+q}(\boldsymbol{\lambda}_*)$ 表示卡方变量加权和 $\sum_{j=1}^{p+q}\lambda_{*j}Z_j^2$ 的 cdf。Z_j^2 是 iid $\chi^2(1)$,而 λ_{*j} 是 $(p+q)\times(p+q)$ 阶矩阵

$$\mathbf{W}=\begin{bmatrix} -\mathbf{B}_f(\boldsymbol{\theta}_*)\mathbf{A}_f(\boldsymbol{\theta}_*)^{-1} & -\mathbf{B}_{fg}(\boldsymbol{\theta}_*,\boldsymbol{\gamma}_*)\mathbf{A}_g(\boldsymbol{\gamma}_*)^{-1} \\ -\mathbf{B}_{fg}(\boldsymbol{\gamma}_*,\boldsymbol{\theta}_*)\mathbf{A}_f(\boldsymbol{\theta}_*)^{-1} & -\mathbf{B}_g(\boldsymbol{\gamma}_*)\mathbf{A}_g(\boldsymbol{\gamma}_*)^{-1} \end{bmatrix} \tag{8.45}$$

的特征值,其中,$\mathbf{A}_f(\boldsymbol{\theta}_*)=\mathrm{E}_0[(\partial^2\ln f/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}')$,$\mathbf{B}_f(\boldsymbol{\theta}_*)=\mathrm{E}_0[(\partial\ln f/\partial\boldsymbol{\theta})(\partial\ln f/\partial\boldsymbol{\theta}')]$,对于密度 $g(\cdot)$,可类似定义矩阵 $\mathbf{A}_g(\boldsymbol{\gamma}_*)$ 与 $\mathbf{B}_g(\boldsymbol{\gamma}_*)$,交叉矩阵 $\mathbf{B}_{fg}(\boldsymbol{\theta}_*,\boldsymbol{\gamma}_*)=\mathrm{E}_0[(\partial\ln f/\partial\boldsymbol{\theta})(\partial\ln g/\partial\boldsymbol{\gamma}')]$,而且期望是关于真实数据生成过程的。对于这些结果的解释与推导,参见翁(Vuong, 1989)。

相反,当 $f(y|\mathbf{x},\boldsymbol{\theta}_*)\neq g(y|\mathbf{x},\boldsymbol{\gamma}_*)$ 时,在 H_0 下有:

$$N^{-1/2}\mathrm{LR}(\hat{\boldsymbol{\theta}},\hat{\gamma})\xrightarrow{d}\mathcal{N}[0,\omega_*^2]\tag{8.46}$$

其中：

$$\omega_*^2=\mathrm{V}_0\left[\ln\frac{f(y|\mathbf{x},\boldsymbol{\theta}_*)}{g(y|\mathbf{x},\boldsymbol{\gamma}_*)}\right]\tag{8.47}$$

而方差是关于真实数据生成过程的。其推导过程，可再次参见翁(Vuong，1989)。

应用这些结果，会随着假定一个模型是否被正确设定以及两个模型之间是否有嵌套关系而变化。

翁对三种模型加以比较并辨别。模型 F_θ 与 G_γ 是：(1)嵌套的(**nested**)，满足 G_γ 嵌套在 F_θ 之中，如果 $G_\gamma\subset F_\theta$；(2)严格非嵌套的(**strictly non-nested**)，当且仅当 $F_\theta\cap G_\gamma=\phi$，因此，两者之中的任一模型都不是另一个模型的特殊化；(3)交叠的(**overlapping**)，如果 $F_\theta\cap G_\gamma\neq\phi$，并且 $F_\theta\not\subseteq G_\gamma$ 以及 $G_\gamma\not\subseteq F_\theta$。佩萨兰和佩萨兰(Pesaran and Pesaran，1995)曾做出一种类似区别。

不论是(2)还是(3)，都是非嵌套模型，但它们需要不同的检验方法。严格非嵌套模型的例子是含有不同误差分布的线性模型，也是含有相同误差分布但对条件均值函数来说不同形式的非线性回归模型。对于交叠模型来说，两个模型的某些特殊化是相等的。一个例子是，含有一些相同的回归元而一些回归元不同的线性模型。

嵌套模型

就嵌套模型而言，一定有 $f(y|\mathbf{x},\boldsymbol{\theta}_*)=g(y|\mathbf{x},\boldsymbol{\gamma}_*)$ 的情况。对于 G_γ 嵌套在 F_θ 之中，H_0 是对 H_f：$\mathrm{E}_0[\ln(f/g)]>0$ 的检验。

一旦利用式(8.45)中 $\mathbf{W}$ 的样本类似形式的特征值 $\hat{\lambda}_j$，对密度可能错误设定来说，加权卡方结果(8.44)是合适的。如若不然，人们能使用如下较小矩阵的样本类似形式的特征值 $\hat{\lambda}_j$：

$$\underline{\mathbf{W}}=\mathbf{B}_f(\boldsymbol{\theta}_*)[\mathbf{D}(\boldsymbol{\gamma}_*)\mathbf{A}_g(\boldsymbol{\gamma}_*)^{-1}\mathbf{D}(\boldsymbol{\gamma}_*)'-\mathbf{A}_f(\boldsymbol{\theta}_*)^{-1}]$$

其中，$\mathbf{D}(\boldsymbol{\gamma}_*)=\partial\phi(\boldsymbol{\gamma}_*)/\partial\boldsymbol{\gamma}$，而约束的准 MLE $\tilde{\boldsymbol{\theta}}=\phi(\hat{\gamma})$，参见翁(Vuong，1989)论文。这一结果提供关于嵌套模型的标准 LR 检验的稳健形式。

当密度 $f(\cdot)$ 确实被正确设定时，或更一般地满足信息矩阵等式，就得到了预期结果：$2\mathrm{LR}(\hat{\boldsymbol{\theta}},\hat{\gamma})\xrightarrow{d}\chi^2(p-q)$，从而 $\mathbf{W}$ 或 $\underline{\mathbf{W}}$ 的特征值 $(p-q)$ 等于 1，而其他情况下则等于 0。

严格非嵌套模型

就严格非嵌套而言，一定有 $f(y|\mathbf{x},\boldsymbol{\theta}_*)\neq g(y|\mathbf{x},\boldsymbol{\gamma}_*)$ 的情况。运用正态分布结果(8.46)，ω_*^2 的一致估计值是：

$$\hat{\omega}^2=\frac{1}{N}\sum_{i=1}^{N}\left(\ln\frac{f(y_i|\mathbf{x}_i,\hat{\boldsymbol{\theta}})}{g(y_i|\mathbf{x}_i,\hat{\gamma})}\right)^2-\left(\frac{1}{N}\sum_{i=1}^{N}\ln\frac{f(y_i|\mathbf{x}_i,\hat{\boldsymbol{\theta}})}{g(y_i|\mathbf{x}_i,\hat{\gamma})}\right)^2\tag{8.48}$$

因而，形成：

$$\mathrm{T}_{\mathrm{LR}}=N^{-1/2}\mathrm{LR}(\hat{\boldsymbol{\theta}},\hat{\gamma})/\hat{\omega}\xrightarrow{d}\mathcal{N}[0,1]\tag{8.49}$$

对临界值为 c 的检验来说，当 $T_{LR}>c$ 时，拒绝 H_0，支持 H_f：$E_0[\ln(f/g)]>0$；当 $T_{LR}<-c$ 时，拒绝 H_0，支持 H_g：$E_0[\ln(f/g)]<0$；而当 $|T_{LR}|<c$ 时，在两个模型之间区分是不可能的。对此检验加以修改，以便允许对数似然惩罚类似于 AIC 与 BIC；参见翁(Vuong, 1989，第 316 页)。与式(8.49)渐近等价的统计量，用恰好等于式(8.48)右边第一项的 $\hat{\omega}^2$ 代替 $\tilde{\omega}^2$。

这种检验假定，两个模型都被错误设定。相反，当假设其中一个模型被正确设定，就要运用 8.5.2 节的考克斯方法。

交叠模型

就交叠模型而言，至于是否有 $f(y|\mathbf{x},\boldsymbol{\theta}_*)=g(y|\mathbf{x},\boldsymbol{\gamma}_*)$ 的情况，其先验信息不清楚，而首先需要人们去检验这个条件。

翁(Vuong, 1989)提出，检验式(8.47)定义的 ω_*^2 是否等于 0，因为 $\omega_*^2=0$ 当且仅当 $f(\cdot)=g(\cdot)$。因此，要计算式(8.48)中的 $\hat{\omega}^2$。在 H_0^{ω}：$\hat{\omega}_*^2=0$ 下，有：

$$N\hat{\omega}^2\xrightarrow{d}M_{p+q}(\boldsymbol{\lambda}_*) \tag{8.50}$$

其中，$M_{p+q}(\boldsymbol{\lambda}_*)$ 的分布已在式(8.44)后面加以定义。利用式(8.45)中 $\mathbf{W}$ 的样本类似形式的特征值 $\hat{\lambda}_j$，当在水平 α 上 $N\hat{\omega}^2$ 大于 $M_{p+q}(\hat{\boldsymbol{\lambda}})$ 分布的 α 百分位数时，就拒绝 H_0^{ω} 假设。否则，更简单地，人们能检验 $\boldsymbol{\theta}_*$ 与 $\boldsymbol{\gamma}_*$ 必须满足 $f(\cdot)=g(\cdot)$ 的条件。为此，利恩和翁(Lien and Vuong, 1987)已经给出一些例子。

倘若不拒绝 H_0^{ω} 或不拒绝 $f(\cdot)=g(\cdot)$ 条件，则其结论是，不可能对给定数据时的两个模型加以区分。若拒绝 H_0^{ω} 或拒绝 $f(\cdot)=g(\cdot)$ 条件，则 H_0 对 H_f 或 H_g 利用 T_{LR} 检验来进行，更详细的内容如同在严格非嵌套的情况下所述。在后一种情况下，显著性水平至多是两个检验中显著性水平最大值的那一个。

这个检验假定两个模型均被错误设定。相反，当假定一个模型被正确设定时，由于两个模型是等价的，另一个模型也必被正确设定。因此，在 H_0 下，$f(y|\mathbf{x},\boldsymbol{\theta}_*)=g(y|\mathbf{x},\boldsymbol{\gamma}_*)$，并能直接利用加权卡方结果(8.44)去变动 LR 检验。设 c_1 与 c_2 分别表示上侧尾部临界值与下侧尾部临界值。当 $2\text{LR}(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\gamma}})>c_1$ 时，拒绝 H_0，支持 H_f；当 $2\text{LR}(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\gamma}})<c_2$ 时，拒绝 H_0，支持 H_g；否则，无法确定检验结果。

8.5.4 *其他非嵌套模型比较*

前面的方法都被限制在完全参数模型上。对仅仅作为部分参数化的模型，例如，不具有正态性假设的线性回归，进行辨别的方法就不太清楚。

8.5.1 节的信息准则能利用损失函数而不是 KLIC 所发展起来的准则来代替。雨宫(Amemiya, 1980)曾经阐述对应于各种不同损失函数的一系列测量。这些测量经常引发嵌套模型，但也可用于非嵌套模型。

一种简单方法是比较预测能力，即选取具有最小均方误差 $(N-q)^{-1}\sum_i(y_i-\hat{y}_i)^2$ 值的那种模型。对线性回归而言，这等价于选择含有最大调整 R^2 的模型，通常认为它提供了很小模型复杂性的惩罚。对非参数模型的一种改进是，去掉一个进行交叉验证(**leave-one-out cross-validation**)(参见 9.5.3 节)。

在非似然情况下，辨别非嵌套模型的正式检验，常常是采用两种方法之一。一

种方法是由戴维森和麦金农(Davidson and MacKinnon, 1984)提出的人工嵌套(**artificial nesting**),该方法把两个嵌套模型嵌套到一个更一般的人工模型之内,从而导致所谓的 J 检验与 P 检验,还有其他一些有关检验。另一种方法是由米宗和理查德(Mizon and Richard, 1986)提出的包容原理(**encompassing principle**),该方法导致一个相当一般的框架,用于检验一个模型与其竞争的一个非嵌套模型。怀特(White, 1994)将此方法与 CM 检验联系起来。对此类文献概述,可参见戴维森和麦金农(Davidson and MacKinnon, 1993,第 11 章)。

8.5.5 *非嵌套模型的例子*

从泊松模型中生成 100 个观测值的样本,该泊松模型均值 $\mathrm{E}[y|\mathbf{x}]=\exp(\beta_1+\beta_2x_2+\beta_3x_3)$,其中,$x_2,x_3\sim\mathcal{N}[0,1]$,$(\beta_1,\beta_2,\beta_3)=(0.5,0.5,0.5)$。因变量 y 的均值为 1.92,标准差为 1.84。两个不正确的非嵌套模型可由泊松回归加以估计:

$$\text{模型 1: } \hat{\mathrm{E}}[y|\mathbf{x}]=\exp(\underset{(8.08)}{0.0608}+\underset{(4.03)}{0.291}x_2)$$

$$\text{模型 2: } \hat{\mathrm{E}}[y|\mathbf{x}]=\exp(\underset{(5.14)}{0.493}+\underset{(5.10)}{0.359}x_3+\underset{(1.78)}{0.091}x_3^2)$$

其中,t 统计量已由下面括号给出。

表 8.2 的前三行给出各种不同的信息准则,具有最小值的模型更好。第一个并没有惩罚参数个数,而且支持模型 2。式(8.40)与式(8.41)中定义的第二个与第三个测量给出了对模型的较大惩罚,具有额外的参数,但仍导致支持较大的模型 2。

表 8.2 泊松回归非嵌套模型比较例子[a]

检验类型	模型 1	模型 2	结论
$-2\ln L$	366.86	352.18	第二个模型更好
AIC	370.86	358.18	第二个模型更好
BIC	376.07	366.00	第二个模型更好
$N\hat{\omega}^2$	以 $p=0.000$ 具有 7.84		能区分
$T_{LR}=N^{-1/2}LR/\hat{\omega}$	以 $p=3.777$ 具有 -0.883		没有模型受到支持

[a] $N=100$。模型 1 是 y 对截距及 x_2 的泊松回归。模型 2 是 y 对截距 x_3 和 x_3^2 的泊松回归。最后两行是对非交叠模型的翁检验(参见正文)。

表 8.2 的最后两行归纳总结了翁检验,即交叠模型的检验。

首先,当在伪真实值处计算时,要对密度等式条件进行检验。已知密度表达式,很容易计算出式(8.48)的统计量 $\hat{\omega}^2$。困难部分是计算式(8.45)中 $\mathbf{W}$ 矩阵的估计值。对泊松密度而言,可能使用 5.2.3 节结尾定义的 $\hat{\mathbf{A}}$ 与 $\hat{\mathbf{B}}$,以及 $\hat{\mathbf{B}}_{fg}=N^{-1}\sum_i(y_i-\hat{\mu}_{fi})\mathbf{x}_{fi}(y_i-\hat{\mu}_{gi})\mathbf{x}'_{gi}$。$\hat{\mathbf{W}}$ 的特征值是 $\lambda_1=0.29$,$\lambda_2=1.00$,$\lambda_3=1.06$,$\lambda_4=1.48$ 以及 $\lambda_5=2.75$。检验统计量 $N\hat{\omega}^2$ 的 p 值具有由式(8.44)给出的分布,该 p 值作为抽取 $\sum_{j=1}^5\lambda_jz_j^2$ 的比例而获得,比如说抽取 10 000 次,这大于 $N\hat{\omega}^2=69.14$。此处,$p=0.000<0.05$,从而我们得出结论:在两个模型之间进行辨别是可能的。在水平 0.05 上,这个例子的临界值等于 16.10,它比 $\chi^2_{0.05}(5)=11.07$ 大许多。

已知区分模型是可行的，那么就能应用第二个检验。此处，$T_{LR}=-0.883$ 支持模型 2，因为它是负的。不过，利用 5%水平上的标准正态双侧检验，其差异并不是统计显著的。在该例子中，$\hat{\omega}^2$ 是相当大的，这意味着第一个检验统计量 $N\hat{\omega}^2$ 是很大的，但第二个检验统计量 $N^{-1/2}\mathrm{LR}(\hat{\boldsymbol{\theta}},\hat{\gamma})/\hat{\omega}$ 是很小的。

8.6 检验结果

在实际应用中，寻找到更好模型之前，要实施一个以上的检验。这会出现实践者时常忽略的几种复杂情况。

8.6.1 *预先检验估计*

为了选取模型，使用设定检验，这样做会使估计量的分布复杂化。例如，假定我们根据在 5%水平上的统计检验，在两个估计量 $\hat{\boldsymbol{\theta}}$ 与 $\tilde{\boldsymbol{\theta}}$ 之间进行选择。比如，$\hat{\boldsymbol{\theta}}$ 与 $\tilde{\boldsymbol{\theta}}$ 可以是无约束模型的估计量与约束模型的估计量。于是，实际估计量是 $\boldsymbol{\theta}^+=w\hat{\boldsymbol{\theta}}+(1-w)\tilde{\boldsymbol{\theta}}$，当检验支持 $\hat{\boldsymbol{\theta}}$ 时，则随机变量 w 取值为 1；而当检验支持 $\hat{\boldsymbol{\theta}}$ 时，则随机变量 w 取值 0。总之，估计量依赖于约束估计量与无约束估计量，并且依赖于随机变量 w，同样也依赖于检验的显著性水平。因此，$\boldsymbol{\theta}^+$ 是一个具有复杂性质的估计量。这称为预先检验估计量（**pretest estimator**），因为该估计量建立在初始检验的基础上。$\boldsymbol{\theta}^+$ 分布可通过正态性下的线性回归模型来获得，并且是非标准的。

从理论上看，统计推断应建立在 $\boldsymbol{\theta}^+$ 分布的基础上。实际应用中，若忽略 w 的随机性，当 $w=1$ 时，将推断建立在 $\hat{\boldsymbol{\theta}}$ 分布的基础上，或当 $w=0$ 时，则将推断建立在 $\tilde{\boldsymbol{\theta}}$ 分布的基础上。为了简单起见就这样做，因为甚至在最简单的模型中，当实施这几类检验时，估计量分布也会变得难以处理。

8.6.2 *检验顺序*

根据实施检验的顺序不同，人们获得各种不同结论。

一种可行的顺序是从一般到特殊（**general to specific**）模型。例如，在对来自消费者需求理论的约束譬如同质性与对称性进行检验之前，人们估计需求的一般模型。或者，整个过程可从特殊到一般（**specific to general**）模型，伴有需要添加回归元及额外的复杂情况，譬如控制内生性，倘若存在，当选择哪一个回归元进入模型时，这种顺序是自然而然的，但当还要实施设定检验时，一种普遍做法是，在同样的一项研究中，既运用一般到特殊的顺序，又运用特殊到一般的顺序。

一个相关问题是，联合检验与单独检验（**joint versus separate tests**）。例如，两个回归元的显著性可以通过两个显著性 t 检验来加以验证，也可以通过联合 F 检验或显著性 $\chi^2(2)$ 检验来验证。一般性讨论已在 7.2.7 节给出，而例子稍后由 18.7 节给出。

8.6.3 *数据挖掘*

作为一种极端形式，广泛运用的选择模型的检验被称为数据挖掘（**data mining**）[洛弗尔（Lovell，1983）]。例如，人们可在几百个可能的 y 的预测元之间进行

探索，然后选取仅仅在双侧检验水平5%上是显著的那些预测元。存在自动搜索的计算机程序，在应用统计学的一些分支上，已被广泛应用。不幸的是，这种广泛搜索将发现伪关系，因为具有水平0.05的检验会产生时间的5%统计显著性的错误发现。洛弗尔指出，应用这种方法倾向于高估拟合优度测量与低估回归系数的抽样方差，甚至当它成功揭露以数据生成过程为特色的变量时。一旦使用标准检验，并报告没有考虑搜索模型程序的 p 值，该方法会使人产生误解，因为名义 p 值与实际 p 值是不同的。怀特(White，2001b)以及沙利文、蒂默曼和怀特(Sullivan，Timmerman，and White，2001)证明，如何运用自助法计算回归元的真实统计显著性。还可参见P.汉森(P. Hansen，2003)。

有时，数据挖掘的动因是为了保存自由度，或避免过度参数化("杂乱")。更重要的是，设定的诸多方面，比如协变量的函数形式，都未能由基本理论解决。已知设定的不确定性，存在判断搜索设定正确的依据[萨根(Sargan，2001)]。不过，当对小样本进行分析，并设定研究的数目相对于样本量很大时，就要格外小心谨慎。当设定研究是时序的，并且有相当多的步骤，同时每一步都要由前面检验结果来确定，该种程序的统计性质总体上是复杂的，且在解析形式上难以处理。

8.6.4 *实用方法*

应用微观经济计量学研究通常运用明确的假设检验来最小化预先检验问题。经济理论用于指导对回归元的选择，大大减少潜在回归元的数目。当样本量很大时，通过去掉"不显著的"变量，使得目标变小。最终结果常常是运用包括用于控制变量的、统计不显著的回归元，诸如工资回归中的地区、行业以及职业等虚拟变量。通过不报告完全模型设定中不重要的系数，能够避免聚集，但要在合适的地方注意这样一种事实。这会导致在估计所关注的重要回归元时，损失一些准确性，但可预防由错误去掉应该被包括进入的变量而引起的偏倚。

一种好的实用做法是，对设定探究及模型选择来说，仅使用部分样本("训练样本")，然后利用完全独立的部分样本("估计样本")，去报告运用偏爱模型所估计出的结果。在这种情况下，倘若子样本是独立的，预先检验就不会影响到估计量分布。由于在最后估计时，利用不完全样本会导致估计量准确性的损失，故该方法通常只有在样本量非常大时才会应用。

8.7 模型诊断

本节讨论非线性模型的拟合优度测量和残差的定义。一种有用的测量是那些在某个特定方面揭示模型不足的测量。

8.7.1 *伪 R^2 测量*

拟合优度被解释为拟合值对因变量样本值的接近程度。

对具有 K 个回归元的线性模型来说，一种最直接的测量是回归标准误差(**standard error of the regression**)，它是误差项的估计标准差：

$$s=\left[\frac{1}{N-K}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2\right]^{1/2}$$

例如，在对数工资回归中，0.10 的回归标准误差，意味着拟合值的大致 95%处于对数工资真实值的 0.20 之内，或处于利用 $e^{0.2}\simeq 1.22$ 的真实工资的 22%之内。除自由度校正之外，这种测量与样本均方根误差是一样的，其中，$\hat{y}_i$ 被看成是 y_i 的预测。作为一种可供选择的方法，人们使用平均绝对误差(**mean absolute error**) $(N-K)^{-1}\sum_i|y_i-\hat{y}_i|$。当非线性模型得出因变量的预测值 $\hat{y}_i$，同样的测算能用于非线性回归模型上。

线性模型中有关的测量是 R^2，即多重测定系数(**coefficient of multiple determination**)。它解释了由回归元解释因变量变异的部分。统计量 R^2 比 s 更广泛地被报告，尽管 s 在计算拟合度时可能包含更有价值的信息。

伪 R^2(**pseudo-R^2**)是 R^2 对非线性回归模型的推广。线性模型的 R^2 有几种解释。这导致了非线性模型中不同的几种可能的伪 R^2 测量，并且不一定具有位于 0～1 之间的性质，以及随添加回归元而增大的性质。为了简单起见，我们阐述不调整自由度的几种此类测量。

一种方法是把 R^2 建立在总平方和(TSS)分解的基础上，得出：

$$\sum_i(y_i-\bar{y})^2=\sum_i(y_i-\hat{y}_i)^2+\sum_i(\hat{y}_i-\bar{y})^2+2\sum_i(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})$$

右边第一项和是残差平方和(RSS)，而第二项是被解释平方和(ESS)。从而，得到两种可行的测量：

$$R^2_{\mathrm{RES}}=1-\mathrm{RSS}/\mathrm{TSS}$$
$$R^2_{\mathrm{EXP}}=\mathrm{ESS}/\mathrm{TSS}$$

对于含有截距线性模型的 OLS 回归来说，第三项和等于 0，所以 $R^2_{\mathrm{RES}}=R^2_{\mathrm{EXP}}$。但是，这种简化在其他模型中不会出现，而且在非线性模型中，通常 $R^2_{\mathrm{RES}}\neq R^2_{\mathrm{EXP}}$。测量 R^2_{RES}能小于 0，R^2_{EXP}能大于 1，同时这两个测量随着添加回归元而减少，尽管对非线性模型的 NLS 回归而言，R^2_{RES}将增大，从而该估计量对 RSS 求极小值。

一种紧密相关的测量是使用：

$$R^2_{\mathrm{COR}}=\widehat{\mathrm{Cor}}^2[y_i,\hat{y}_i]$$

即真实值与拟合值之间的平方相关系数。测量 R^2_{COR}位于 0～1 之间，对于含有截距的线性模型来说，它等于 OLS 回归的 R^2。在非线性模型中，R^2_{COR}随着添加回归元增多而递减。

第三种方法是，使用加权平方和，以便控制横截面数据的内在异方差性。设 $\hat{\sigma}_i^2$ 表示 y_i 的拟合条件方差，其中，假定异方差性可以用显性方式进行建模，对于 FGLS 以及一些模型譬如 logit 和泊松模型来说，确实如此。那么，我们能使用：

$$R^2_{\mathrm{WSS}}=1-\mathrm{WRSS}/\mathrm{WTSS}$$

其中，加权残差平方和 $\mathrm{WRSS}=\sum_i(y_i-\hat{y}_i)/\hat{\sigma}_i^2$，$\mathrm{WTSS}=\sum_i(y_i-\hat{\mu})/\hat{\sigma}^2$，而 $\hat{\mu}$ 与 $\hat{\sigma}^2$

分别表示仅有截距的模型中的估计均值与方差。这被称为皮尔逊 R^2，因为 WRSS 等于皮尔逊统计量，除任何有限样本校正之外，若异方差性被正确建模，这应等于 N。注意到，R^2_{WSS} 可以小于 0，且随着添加回归元增多而减少。

第四种方法是，R^2 针对目标函数而不是平方残差和加以推广。设 $Q_N(\boldsymbol{\theta})$ 表示被求极大值的目标函数，Q_0 表示仅含有截距模型时的值，Q_{fit} 表示拟合模型的值，而 $Q_{\max}$ 表示 $Q_N(\boldsymbol{\theta})$ 的最大可能值。于是，因包含回归元而引起的目标函数的最大潜在收益是 $Q_{\max}-Q_0$，而实际收益是 $Q_{\mathrm{fit}}-Q_0$。这就建议使用：

$$R^2_{\mathrm{RG}}=\frac{Q_{\mathrm{fit}}-Q_0}{Q_{\max}-Q_0}=1-\frac{Q_{\max}-Q_{\mathrm{fit}}}{Q_{\max}-Q_0}$$

进行测量，其中，下标 RG 意味着相对收益(**relative gain**)。就最小二乘估计而言，损失函数极大化是负的残差平方和。从而，$Q_0=-\mathrm{TSS}$，$Q_{\mathrm{fit}}=-\mathrm{RSS}$ 以及 $Q_{\max}=0$，因此，对 OLS 或 NLS 回归来说，$R^2_{\mathrm{RG}}=\mathrm{ESS}/\mathrm{TSS}$。测量 R^2_{RG} 具有位于 0～1 之间的优点，且随着添加回归元增多而增大。就极大似然估计而言，损失函数是 $Q_N(\boldsymbol{\theta})=\ln L_N(\boldsymbol{\theta})$。于是，不能总是使用 R^2_{RG}，因为在一些模型中，$Q_{\max}$ 可能是无界的。例如，对线性模型来说，在正态性下，当 $\sigma^2\to 0$ 时，$L_N(\boldsymbol{\beta},\sigma^2)\to\infty$。对线性指数族模型的极大似然以及拟极大似然估计来说，譬如 logit 与泊松，$Q_{\max}$ 通常是已知的，并且可以证明，R^2_{RG} 是建立在下一节定义的残差离差的基础上。

与 R^2_{RG} 有关的测量是 $R^2_Q=1-Q_{\mathrm{fit}}/Q_0$。该测量会随着添加回归元增多而增大。当 $Q_{\max}=0$ 时，它等于 R^2_{RG}，这正是 OLS 回归与二值及多项式模型的情况。另外，就离散数据而言，这一测量可能具有小于 1 的上界；而对连续数据来说，此测量可能并不介于 0 与 1 之间，这是因为，对数似然可以是负的或正的。例如，对具有连续密度的 ML 估计来说，可能出现 $Q_0=1$ 且 $Q_{\mathrm{fit}}=4$，导致 $R^2_Q=-3$；或者可能出现 $Q_0=-1$ 且 $Q_{\mathrm{fit}}=4$，导致 $R^2_Q=5$。

因此，对非线性模型来说，不存在普遍性的伪 R^2。最有用的测量或许是 R^2_{COR}，因为相关系数容易进行解释，以及在 $Q_{\max}$ 为已知的一些特殊情况下的 R^2_{RG}。卡梅伦和温德梅杰(Cameron and Windmeijer, 1997)对许多测量进行了分析，并且卡梅伦和温德梅杰(Cameron and Windmeijer, 1996)则将这些测量用到计数数据模型上。

8.7.2 *残差分析*

与统计学的一些其他领域相比，微观经济计量学分析确实较少强调残差分析。当数据集很小时，关注的内容是，残差分析可能导致对模型的过度拟合。当数据集很大时，就有如下看法，没有必要进行残差分析，因为单个观测值对分析具有很小的影响。因此，我们给出一个简要的综述。例如，更全面的讨论已由麦卡拉和内尔德(McCullagh and Nelder, 1989)以及卡梅伦和特里维迪(Cameron and Trivedi, 1998，第 5 章)给出。特别地，经济计量学家对删失模型和截取模型中定义的残差感兴趣。

就非线性回归模型而言，已提出一系列范围广泛的残差。考察纯量因变量 y_i，其拟合值 $\hat{y}_i=\hat{\mu}_i=\mu(\mathbf{x}_i,\hat{\boldsymbol{\theta}})$。原始残差(**raw residual**)是 $r_i=y_i-\hat{\mu}_i$。皮尔逊残差

(**Pearson residual**)是对异方差性 $p_i=(y_i-\hat{\mu}_i)/\hat{\sigma}_i$ 的一种明显修正，其中，$\hat{\sigma}_i$ 表示 y_i 的条件方差的估计值。这就需要对 y_i 方差进行设定，对泊松模型就是这样做的。对 LEF 密度来说(参见 5.7.3 节)，离差残差(**deviance residual**)是 $d_i=\text{sign}(y_i-\hat{\mu}_i)\sqrt{2[l(y_i)-l(\hat{\mu}_i)]}$，其中，$l(y)$表示 $y|\mu$ 的对数密度在 $\mu=y$ 处的计算值，而 $l(\hat{\mu})$表示在 $\mu=\hat{\mu}$ 处的计算值。离散残差的动因是，这些残差的平方和作为离差统计量，它是对线性模型中的原始残差和的 LEF 模型的一种推广。安斯科姆残差(**Anscombe residual**)被定义成 y 的一种变换，该变换使得 y 最接近于正态性，然后正规化成均值为 0 且方差为 1 的情况。就 LEF 密度而言，可获得这样的变换。

为了解释 $\hat{\mu}_i$ 中的估计误差，提出一种残差对小样本进行修正。对线性模型而言，这需要残差被 $\sqrt{1-h_{ii}}$ 去除，其中，h_{ii}表示帽矩阵 $\mathbf{H}=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}$ 中的第 i 个对角元素。这些残差被认为具有较好的有限样本特性。由于 $\mathbf{H}$ 的秩为 K，即回归元的个数，h_{ii}的平均值是 K/N，而且大于 $2K/N$ 的 h_{ii}的一些值被看成具有高的杠杆作用。这些结果可扩展到满足 $\mathbf{H}=\mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}\mathbf{W}^{1/2}$ 的 LEF 模型上，这里，$\mathbf{W}=\text{Diag}[w_{ii}]$，$w_{ii}=g'(\mathbf{x}_i'\boldsymbol{\beta})/\sigma_i^2$，而 $g(\mathbf{x}_i'\boldsymbol{\beta})$ 与 σ_i^2 分别表示设定的条件均值与方差。麦卡拉和内尔德(McCullagh and Nelder, 1989)曾提供一个综述。

更一般地讲，考克斯和斯内尔(Cox and Snell, 1968)将广义残差(**generalized residual**)定义为，满足相对弱条件的任何纯量函数 $r_i=r(y_i,\mathbf{x}_i,\hat{\boldsymbol{\theta}})$。得到该残差的一种方式是，许多估计量拥有 $\sum_i\mathbf{g}(x_i,\boldsymbol{\theta})r(y_i,\mathbf{x}_i,\hat{\boldsymbol{\theta}})=\mathbf{0}$ 形式的一阶条件，这里，y_i 出现在纯量 $r(\cdot)$中，但不出现在向量 $g(\cdot)$之中。还可参见怀特(White, 1994)。

对建立在正态潜变量基础上的回归模型(参见第 14 章和第 16 章)来说，切舍和艾里什(Chesher and Irish, 1987)提出，利用 $\mathrm{E}[\varepsilon_i^*|y_i]$作为残差，其中，$y_i^*=\mu+\varepsilon_i^*$ 表示未观测潜变量，而 $y_i=g(y_i^*)$表示观测因变量。对 $g(\cdot)$的一些特殊选取，对应于 probit 模型与 Tobit 模型。古里耶克斯等人(Gouriéroux et al., 1987)把这种方法推广到 LEF 密度上。在此背景下，一种正常方法是，沿着 10.3 节中期望最大算法的线索，把残差看成缺失数据。

对残差的一种普遍使用是，将其用于绘制对其他关注变量的曲线图。残差对拟合值的曲线图能揭示不好的模型拟合；残差对省略变量的曲线图建议包括更多的回归元模型；残差对已包含回归元的曲线图能建立需要不同的函数形式。在这类曲线图中，包括非参数回归线是有益的(参见第 9 章)。当数据只取几个离散值，很难对曲线图给予解释，因为仅在几个值上有聚集，对此使用所谓的不稳定特殊性，即向数据中添加一些随机噪声来减少聚集，则是有益的。

某些参数模型蕴含，适当定义的残差应是正态分布的。这能通过正态分值曲线图来加以检验，即把残差 r_i 从小到大进行排序，若残差确实显示出正态分布，就绘制残差对预测值的曲线图。因此，绘制有序的 r_i 对 $\bar{r}+s_r\boldsymbol{\Phi}^{-1}((i-0.5)/N)$的曲线图，其中，$\bar{r}$ 与 s_r 分别表示 r 的样本均值与标准差，而 $\boldsymbol{\Phi}^{-1}(\cdot)$表示标准正态 cdf 的反函数。

8.7.3 *诊断例子*

表 8.3 使用与 8.5.5 节一样的数据生成过程。因变量 y 具有均值 1.92 和标准差 1.84。y 对 x_3 的泊松回归以及 y 对 x_3 及 x_3^2 的泊松回归是：

$$\text{模型 1: } \hat{\mathrm{E}}[y|\mathbf{x}]=\exp(\underset{(5.20)}{0.586}+\underset{(7.60)}{0.389}x_3)$$

$$\text{模型 2: } \hat{\mathrm{E}}[y|\mathbf{x}]=\exp(\underset{(5.14)}{0.493}+\underset{(5.10)}{0.359}x_3+\underset{(1.78)}{0.091}x_3^2)$$

其中，t 统计量已在括号中给出。

表 8.3 伪 R^2：泊松回归例子[a]

诊 断	模型 1	模型 2	差异
s，其中 $s^2=\mathrm{RSS}/(\mathrm{N}\text{-}\mathrm{K})$	0.166 2	0.166 1	0.000 1
$R^2_{\mathrm{RES}}=1-\mathrm{RSS}/\mathrm{TSS}$	0.188 5	0.196 2	+0.007 7
$R^2_{\mathrm{EXP}}=\mathrm{TSS}/\mathrm{ESS}$	0.166 7	0.208 7	+0.040 2
$R^2_{\mathrm{COR}}=\widehat{\mathrm{Cor}}^2[y_i,\hat{y}_i]$	0.189 3	0.196 4	+0.006 7
$R^2_{\mathrm{WSS}}=1-\mathrm{WRSS}/\mathrm{WTSS}$	0.156 2	0.169 5	+0.023 3
$R^2_{\mathrm{RG}}=(Q_{\mathrm{fit}}-Q_0)/(Q_{\max}-Q_0)$	0.155 2	0.171 2	+0.016 0
$R^2_{\mathrm{Q}}=1-Q_{\mathrm{fit}}/Q_0$	0.077 3	0.080 8	+0.007 5

[a] $N=100$。模型 1 表示 y 对截距与 x_3 的泊松回归。模型 2 表示 y 对截距、x_3 以及 x_3^2 的泊松回归。RSS 表示残差平方和(SS)，ESS 表示被解释平方和(SS)，TSS 表示总平方和，WRSS 表示加权 RSS，WTSS 表示加权 TSS，Q_{fit}表示目标函数的拟合值，Q_0表示仅有截距模型的拟合值，$Q_{\max}$表示给定数据以及仅当某些目标函数存在时目标函数的最大可能值。

在这个例子中，所有的 R^2 测量都随着添加 x_3^2 作为回归元而增大，尽管该例子中，最后的 R^2 几乎具有相似值，但其他的 R^2 却具有截然不同的数值。更一般地讲，前三个 R^2 是标度相似的，而 R^2_{RES} 及 R^2_{COR} 表现十分接近，只是剩下的三个测量却具有十分不同的标度。只有最后两个测量 R^2 才会保证随着添加回归元增多而增大，除非目标函数是平方误差和。这里，可以建立测量 R^2_{RG}，因为如果拟合均值 $\hat{\mu}_i=y_i$，对于所有 i，那么泊松对数似然是极大化的，从而得出 $Q_{\max}=\sum_i[\, y_i\ln y_i-y_i-\ln y_i!\,]$，其中，当 $y=0$ 时，有 $y\ln y=0$。

此外，可计算模型 2 的三个残差。原始残差、皮尔逊残差以及离差残差的样本均值与标准差分别为 0 与 1.65、0.01 与 1.97、−0.21 与 1.22。对原始残差来说，该残差具有 0 均值，是含有截距泊松回归的一个性质，此种性质仅与极少数的其他模型所共有。较大的原始残差标准差，反映出缺乏标度与 y 的标准差大于 1 的事实。这些残差两两之间的相关系数全都大于 0.96。当 R^2 很小时，可能出现这类情况，因此 $\hat{y}_i\approx\bar{y}$ 。

8.8 应用研究

通过运用辅助回归，m 检验与豪斯曼检验都最容易实施。人们应该发现，所做的这些辅助回归假设，只在一些分布假设下才是有效的，这些分布假设比为获得回归系数的通常稳健的标准误差而做出的那些假设要强。一些稳健检验已在 8.4 节阐述。

除了在不现实情况下，即模型的所有方面——函数形式、回归元和分布——都

被正确设定的情况下，对于充分大的数据集以及固定显著性水平，比如 5%，将会拒绝由模型所蕴含的样本矩条件。在经典检验情形下，这经常是人们所希望的结果。特别地，对充分大的样本来说，回归系数将总是显著地异于 0，许多研究都探讨过这类结果。不过，对设定检验而言，人们的要求通常是不被拒绝，因此，人们可以认为，模型是被正确设定的。或许正是因为这个缘由，设定检验才未被充分运用。

举一个例子，考察对消费生命周期模型的正确设定进行检验。除非样本是小的，所研究的设定检验式可能在 5%水平上拒绝模型。例如，假定模型设定检验统计量服从 $\chi^2(12)$分布，当样本量 $N=3\,000$ 时，它的 p 值为 0.02。即使模型在 5%的显著性水平上被拒绝，但生命周期模型却对数据给出了不好的解释，这一点并不清楚。一种可能性是增大临界值，因为一旦利用 BIC（参见 8.5.1 节），样本量将增加。

设定检验未充分使用的另一种原因是，当利用更方便的辅助回归实施检验的渐近等价形式时，会出现计算上的困难与不好的检验水平性质。通过运用自助法，这些缺点将大大减少。第 11 章将阐述自助法，以便实施本章给出的一些检验。

8.9 文献注释

8.2 归功于纽韦（Newey，1985）与陶亨（Tauchen，1985）的条件矩检验，是对怀特（White，1982）的信息矩阵检验的推广。就 ML 估计而言，m 检验通过辅助回归的计算是对 IM 检验的兰开斯特（Lancaster，1984）与切舍（Chesher，1984）方法的推广。帕甘和维拉（Pagan and Vella，1989）给出 m 检验一个很好的概述。m 检验提供一种非常一般的评述检验框架。可以证明，它嵌套所有的检验，譬如沃尔德检验、LM、LR 以及豪斯曼检验。怀特（White，1994）曾强调这种统一性质。

8.3 豪斯曼检验是由豪斯曼（Hausman，1978）提出的，较早的参考文献已由 8.3 节给出，而鲁德（Ruud，1984）给出了一个良好的综述。

8.4 由格林（Greene，2003）、戴维森和麦金农（Davidson and McKinnon，1993），以及伍德里奇（Wooldridge，2002）所撰写的书，都阐述了许多标准的设定检验。

8.5 佩萨兰和佩萨兰（Pesaran and Pesaran，1993）已经讨论过，当得不到对数似然的期望解析表达式时，考克斯非嵌套检验如何得以实施。也可使用翁（Vuong，1989）检验。

8.7 关于非线性模型的模型诊断，往往通过把线性回归模型的结果扩展到广义线性模型譬如 logit 与泊松模型而获得。卡梅伦和特里维迪（Cameron and Trivedi，1998，第 5 章）已经给出详细讨论和相关参考文献。

习 题

8-1 假定 $y=\mathbf{x}'\boldsymbol{\beta}+u$，其中，$u\sim\mathcal{N}[0,\sigma^2]$，参数向量 $\boldsymbol{\theta}=[\boldsymbol{\beta}',\sigma^2]'$，并且密度 $f(y|\theta)=(1/\sqrt{2\pi}\sigma)\exp[-(y-\mathbf{x}'\boldsymbol{\beta})^2/2\sigma^2]$。存在 N 个独立观测值的样本。

(a) 请解释矩条件 $E[\mathbf{x}(y-\mathbf{x}'\beta)^3]$的检验为什么是正态分布误差假设的检验。

(b) 提供式(8.5)给出的 $\hat{\mathbf{m}}_i$ 与 $\hat{\mathbf{s}}_i$ 的表达式时,必须实施建立在(a)部分条件矩基础上的 m 检验。

(c) 假定 $\dim[\mathbf{x}]=10$,$N=100$,并且式(8.5)中的辅助回归会产生 0.2 的非中心化 R^2。你在水平 0.05 上,会得出什么结论呢?

(d) 对这个例子,给出由怀特信息矩阵检验所进行的检验矩条件。

8-2 考察式(8.23)给出的 PCGF 检验的多项式形式,这里,用 $\hat{p}_j=N^{-1}\times\sum_i F(\mathbf{x}_i,\hat{\boldsymbol{\theta}})$代替 p_j。证明,PCGF 能表示成式(8.27)的 CGF,$\hat{\mathbf{V}}=\mathrm{Diag}[N\hat{p}_j]$。(从而得出结论:在多项式情况下,安得鲁斯检验统计量简化为皮尔逊统计量。)

8-3 [改编自雨宫(Amemiya, 1985)。]对于 8.4.1 节给出的豪斯曼检验,设 $V_{11}=V[\hat{\boldsymbol{\theta}}]$,$V_{22}=V[\tilde{\boldsymbol{\theta}}]$以及 $V_{12}=\mathrm{Cov}[\hat{\boldsymbol{\theta}},\tilde{\boldsymbol{\theta}}]$。

(a) 证明估计量 $\bar{\boldsymbol{\theta}}=\hat{\boldsymbol{\theta}}+[V_{11}+V_{22}-2V_{12}]^{-1}(\tilde{\boldsymbol{\theta}},\hat{\boldsymbol{\theta}})$,具有渐近方差矩阵 $V[\bar{\boldsymbol{\theta}}]=V_{11}-[V_{11}-V_{12}][V_{11}+V_{22}-2V_{12}]^{-1}[V_{11}-V_{12}]$。

(b) 证明 $V[\bar{\boldsymbol{\theta}}]$ 在矩阵意义下小于 $V[\hat{\boldsymbol{\theta}}]$,除非 $\mathrm{Cov}[\hat{\boldsymbol{\theta}},\tilde{\boldsymbol{\theta}}]=V[\hat{\boldsymbol{\theta}}]$。

(c) 现在假定 $\hat{\boldsymbol{\theta}}$ 是完全有效的。$V[\bar{\boldsymbol{\theta}}]$ 能小于 $V[\hat{\boldsymbol{\theta}}]$吗?你会得出什么结论呢?

8-4 假定两个模型都是非嵌套的,并存在 $N=200$ 个观测值。对第一个模型来说,参数个数 $q=10$ 且 $\ln L=-400$。对第二个模型来说,$q=10$ 且 $\ln L=-380$。

(a) 哪一个模型有利于利用 AIC?

(b) 哪一个模型有利于利用 BIC?

(c) 若两个模型确实是嵌套的,且在水平 0.05 上使用似然比检验,哪一个模型受到支持呢?

8-5 使用 16.6 节的健康开销支出数据。模型是 DMED 的 probit 回归,即良好健康开销支出的标示变量,对应于 16.6 节第二段中列出的 17 个回归元。你应该求出表 16.1 第 1 列给出的估计值。

(a) 利用豪斯曼检验,在水平 0.05 上对自测健康标示 HLTHG、HLTHF 以及 HLTHP 的联合统计显著性进行检验。(若用软件包计算,可能需要某种额外的编程。)

(b) 此处的豪斯曼检验是最佳检验吗?

(c) 在水平 0.05 上,信息矩阵检验会支持这个模型的约束吗?(这将需要某种额外的编程。)

(d) 根据 R^2_{RES}、R^2_{EXP}、R^2_{COR}和 R^2_{RG},区分去掉 HLTHG、HLTHF、HLTHP 的模型与去掉 IC、IDP、LPI 的模型。

9

半参数方法

9.1 引　论

本章阐述数据分析方法，与前几章方法所需的模型设定相比，本章方法所需的模型设定要少一些。

我们以非参数估计开始。关于数据生成过程，这里做出非常少的假设。一个重要例子是利用核密度估计方法对连续密度进行估计。这因为提供了比熟知的直方图更光滑的形式而引人注目。第二个重要例子是，对纯量回归元进行非参数回归，诸如核回归。这对(x, y)散点图设置了灵活曲线，而没有用曲线形式的参数约束。非参数估计有大量应用，包括数据描述、对来自回归模型的数据和拟合残差进行探索性分析，以及对由蒙特卡罗研究获得的参数估计进行各种不同模拟的概括。

经济计量分析强调，纯量 y 对回归元向量 $\mathbf{x}$ 进行多变量回归。然而，非参数方法尽管在理论上可能具有无限大样本，但在实际应用中却不尽如人意，因为这需要在几个方面对数据加以分切，从而导致在每一个切片中具有极少的数据。

因此，经济计量学家关注于半参数方法。这些方法把非参数成分与参数成分结合起来，从而大大减少了维数。一个重要应用是，允许关于条件均值的更灵活的模型。例如，条件均值 $\mathrm{E}[y|\mathbf{x}]$ 被参数化为单指标形式 $g(\mathbf{x}'\boldsymbol{\beta})$，其中，不用对 $g(\cdot)$ 函数形式设定，却可用非参数形式加以估计，未知参数 $\boldsymbol{\beta}$ 也以非参数形式进行估计。另一个应用是，当出现错误设定分布假设会导致非一致参数估计值时，则运用非参数方法对那些分布假设进行放松。例如，当 y 数据被截取或删失时（参见第16章），假如没有对误差项特定分布做出正确设定，我们希望获得线性回归模型 $y=\mathbf{x}'\boldsymbol{\beta}+\varepsilon$ 中 $\boldsymbol{\beta}$ 的一致估计值。

非参数方法的渐近理论不同于大部分参数方法的那些渐近理论。当 $N\rightarrow\infty$ 时，估计可通过把数据分切成甚至更小的切片，然后在每一个切片中估计局部特性。由于把小于 N 个观测值用于估计每个切片，所以收敛速度慢于前面几章中曾获得的收敛速度。不过，在最简单情况下，非参数估计仍旧服从渐近正态分布。在一些半参数回归的重要情况下，参数 $\boldsymbol{\beta}$估计量拥有以速度 $N^{-1/2}$收敛的通常性质。因此，通过标度$\sqrt{N}$会导致极限正态分布，而此模型的非参数成分却以较慢的速度 N^{-r}收敛，其中，$r<1/2$。

由于非参数方法是局部平均方法，对局部的不同选择会导致各种不同的有限样本结果。在一些约束情况下，存在一些规则和方法来确定用于局部平均的带宽或窗口宽度，正如存在用于确定给定观测值个数时直方图中箱子(bins)个数的规则一样。此外，一种普遍做法是，运用非科学方法选取带宽，即画出看起来合理的光滑图，就能够获捕到所关注关系的详情。

非参数方法构成本章主体，因为非参数方法既是关注的内在内容，也是进入半参数方法的基础，本章尤其要阐述离散因变量与删失因变量模型。这里强调核方法，原因在于核方法阐述相对简单，同时"声称所有光滑方法在渐近意义下，本质上等价于核光滑"[哈德尔(Härdle, 1990, 第 11 页)]。

9.2 节提供非参数密度估计和非参数回归应用到数据上的一些例子。9.3 节阐述核密度估计。局部回归在 9.4 节加以讨论，以便为 9.5 节给出核回归的正式研究提供动因。9.6 节阐述不同于核方法的非参数回归方法。半参数回归的大部分专题则在 9.7 节引入。

9.2 非参数例子：小时工资

举一个例子，我们考察在 1993 年工作、年龄为 36 岁的 175 名妇女的小时工资与受教育情况。数据来自密歇根收入动态面板调查。小时工资的分布是向右偏斜的，这一点很容易建立起来，故将其建模成 ln wage，即小时工资的自然对数。

我们只给出非参数密度估计的一个例子和非参数回归的一个例子，并阐述选择带宽的重要作用。然后，9.3 节给出基本理论。

9.2.1 *非参数密度估计*

工资自然对数的直方图，如图 9.1 所示。为了提供详情，对箱子宽度加以选取，以使存在 30 个箱子，每个箱子宽度为 0.20 左右。仅就 175 个观测值而言，这是异常窄的，但对较大箱子宽度而言，则会损失更多详细信息。工资对数数据似乎是对称的，尽管它们可能稍微向左偏斜。

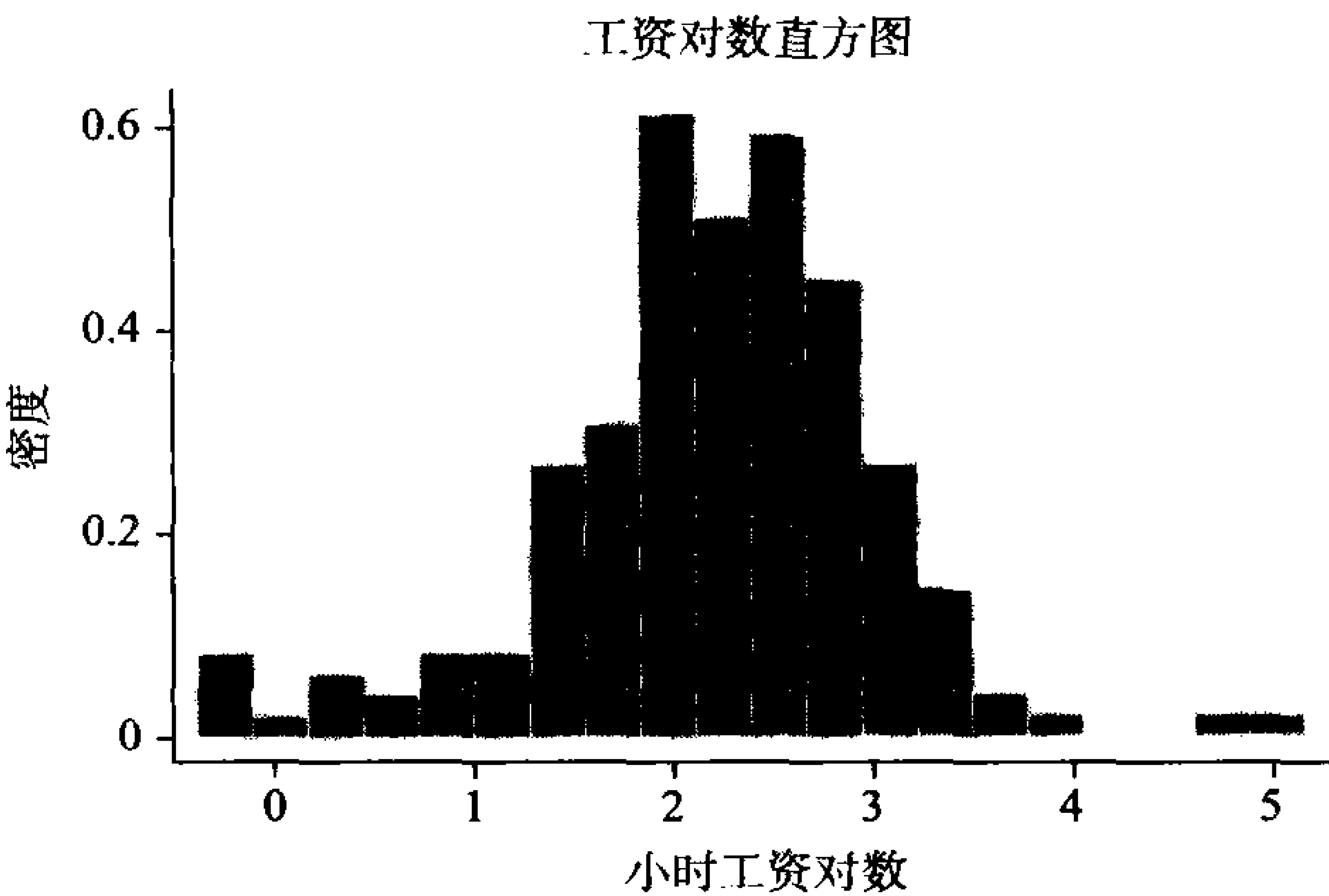


图 9.1 小时工资自然对数的直方图。数据来自美国在 1993 年工作、年龄为 36 岁的 175 名妇女。

标准光滑的非参数密度估计是指由式(9.3)定义的核密度估计。此处，我们使用表 9.1 中由埃帕内尼科夫(Epanechnikov)定义的核。

表 9.1 核函数：普遍使用的例子[a]

核	和函数	δ
一致(或盒形或矩形)	$\frac{1}{2}\times\mathbf{1}(\|z\|<1)$	1.351 0
三角形的(或三角形)	$(1-\|z\|)\times\mathbf{1}(\|z\|<1)$	—
埃帕内尼科夫的(或二次的)	$\frac{3}{4}(1-z^2)\times\mathbf{1}(\|z\|<1)$	1.718 8
四次的(或双权的)	$\frac{15}{16}(1-z^2)^2\times\mathbf{1}(\|z\|<1)$	2.036 2
三次权重的	$\frac{35}{32}(1-z^2)^3\times\mathbf{1}(\|z\|<1)$	2.312 2
三次立方的	$\frac{70}{81}(1-\|z\|^3)^3\times\mathbf{1}(\|z\|<1)$	—
高斯的(或正态的)	$(2\pi)^{-1/2}\exp(-z^2/2)$	0.776 4
四阶高斯的	$\frac{1}{2}(3-z)^2(2\pi)^{-1/2}\exp(-z^2/2)$	—
四阶四次的	$\frac{15}{32}(3-10z^2+7z^4)\times\mathbf{1}(\|z\|<1)$	—

[a] 常值 δ 由式(9.11)定义，并用于获得由式(9.13)给出的西尔弗曼插入估计值。

具体实施时，其根本决策是选择带宽。在这个例子中，式(9.13)定义的西尔弗曼插入估计会产生 $h=0.545$ 的带宽。于是，核估计是下面那些观测值的加权平均，即在当前计算点上对数工资的 0.545[1] 单位内具有对数工资的，并且其最大权数被设置在最靠近当前计算点上。图 9.2 给出带宽分别为 0.273、0.545 以及 1.091 的三个核密度估计，它们分别对应于半个插入、一个插入、2 倍插入的插入带宽。很明显，最小带宽表现得太小，因为它会导致密度估计值凹凸不平。最大带宽使得数据又过分光滑。中间带宽即插入 0.545 值，看起来是一个最佳选择。它给出了合理的光滑密度估计值。

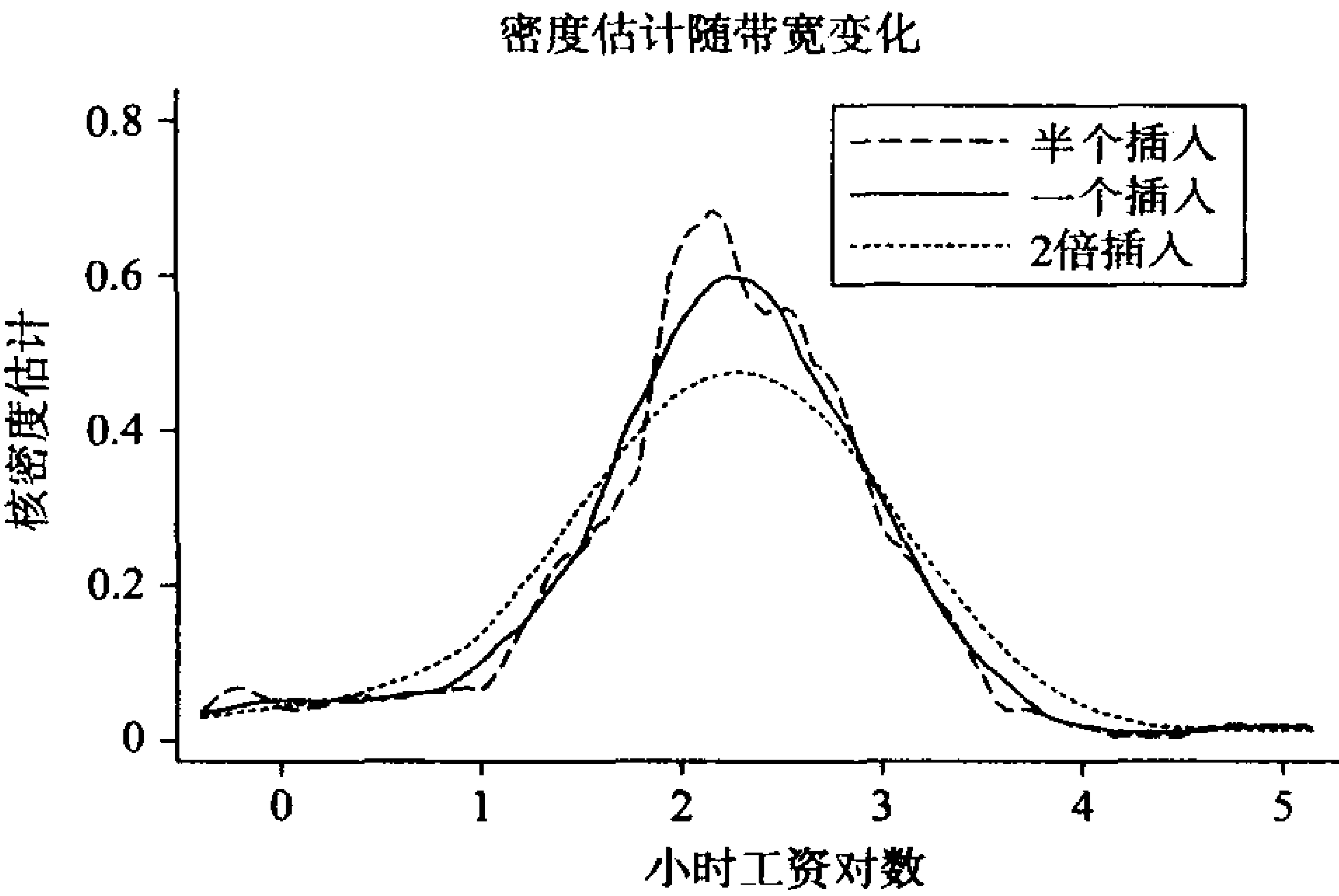


图 9.2 利用埃帕内尼科夫核，对于三种不同带宽的工资对数的核密度估计。

[1] 原著中此处为 0.21，但应为 0.545，这里已改。——译者注

我们利用这个核密度估计会做些什么呢？一种可能是，通过附加均值等于样本均值且方差等于样本方差的正态密度，将该密度与正态的情况加以对比。虽然图形没有重新画出，可是与正态情况相比，此处却揭示出具有更好带宽 0.545 的核密度估计具有更加尖锐的尖峰。第二种可能性是，比较不同子分组譬如受教育程度或全日制或半日制工作状况时的核密度估计。

9.2.2 *非参数回归*

我们考察工资对数与受教育之间的关系。此处所用的非参数方法是洛斯(Lowess)的局部回归方法，即局部加权平均估计量[参见式(9.16)与 9.6.2 节]。

局部加权回归线在每一个点 x 处均利用那种中心子集来拟合，该中心子集包括最接近的 0.8N 个观测值，这是一种默认程序，其中，N 表示样本量，并其权重则随距离 x 越远而下降。对于靠近端点的 x 值来说，使用较小的非中心子集。

图 9.3 给出对数工资与受教育的散点图，以及带宽分别为 0.8、0.4 以及 0.1 的洛斯回归曲线。前两个带宽的散点图给出了类似曲线。其关系表现出二次形式的，但这或许是一种推测，因为在受教育很少的水平上，数据相对稀少。就大部分数据而言，线性关系或许同样能很好地发挥作用。为了简单起见，我们没有阐述 95%的置信区间或前面曾提供的带宽。

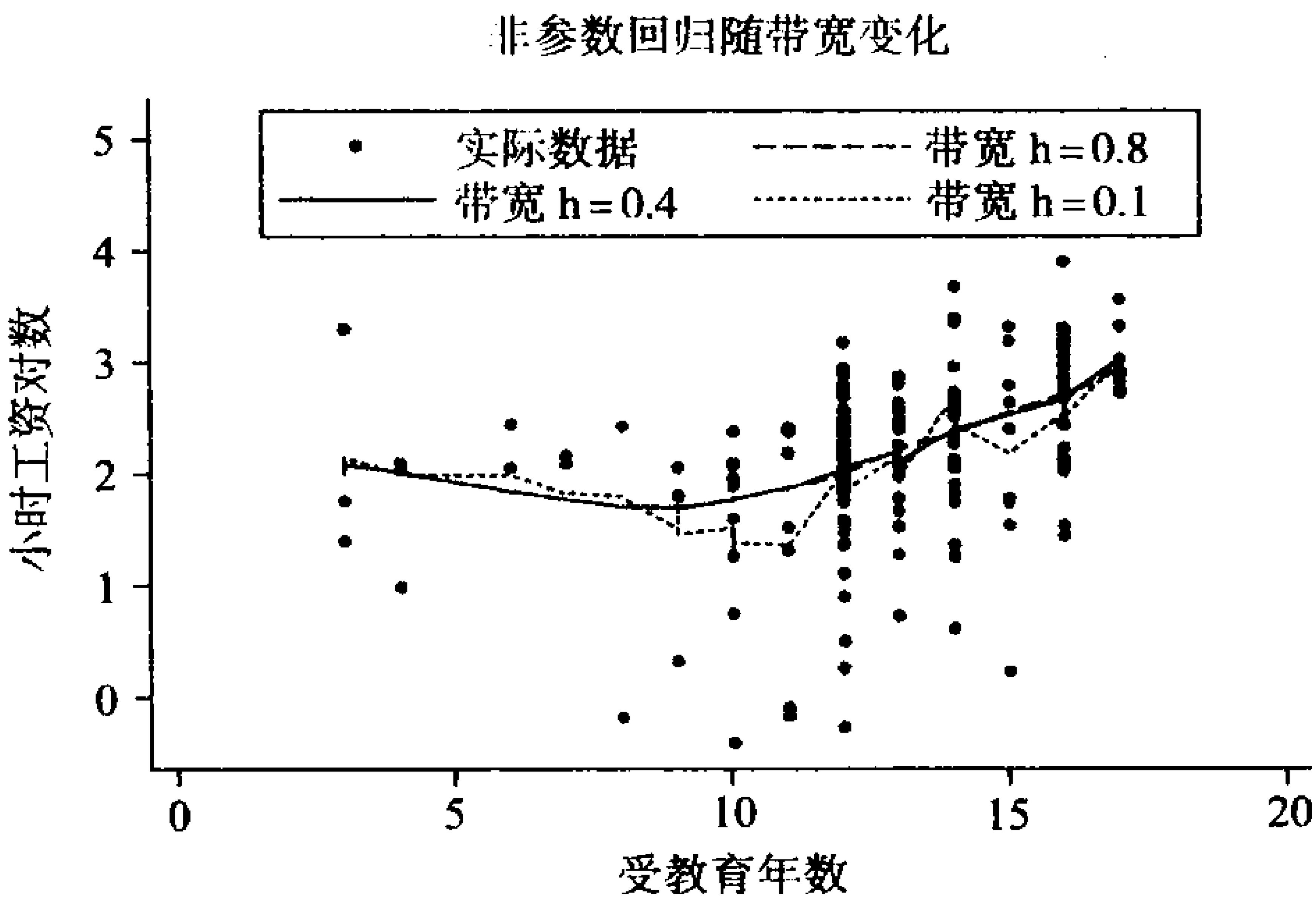


图 9.3 利用洛斯回归，三种不同带宽的工资对数对受教育的非参数估计。样本与图 9.1 的一样。

9.3 核密度估计

比较各种不同分组或与基准密度譬如正态密度相比较时，非参数密度估计十分有用。与直方图比较，非参数密度估计具有提供比较光滑密度估计的优点。与直方图中选取箱子个数相似，其重要决策是选取带宽。我们关注于标准非参数密度估计量，即核密度估计。一旦得出核密度估计的详细描述，则与回归有关的结果可通过运用更为简单的密度估计来得到。

9.3.1 直方图

直方图(**histogram**)是将 x 的范围分割成相等区间的空间,并计算每一个区间上的样本部分值,以此得出密度估计值。

我们给出更正式的直方图表述,这样做会自然地推广到更光滑的核密度估计量上。考察纯量连续随机变量 x 在 x_0 处计算的密度 $f(x_0)$ 的估计。由于密度是 cdf $F(x_0)$ 的导数[也就是说,$f(x_0)=dF(x_0)/dx$],故有:

$$
\begin{aligned}
f(x_0) &= \lim_{h\to 0}\frac{F(x_0+h)-F(x_0-h)}{2h} \\
&= \lim_{h\to 0}\frac{\Pr[x_0-h<x<x_0+h]}{2h}
\end{aligned}
$$

对于样本量为 N 的样本$\{x_i,\ i=1,\cdots,N\}$来说,建议利用估计量:

$$
\hat{f}_{\mathrm{HIST}}(x_0)=\frac{1}{N}\sum_{i=1}^{N}\frac{\mathbf{1}(x_0-h<x_i<x_0+h)}{2h} \tag{9.1}
$$

其中,指示函数[1](indicator function)为:

$$
\mathbf{1}(A)=\begin{cases}1, & \text{若事件 } A \text{ 发生}\\ 0, & \text{其他}\end{cases}
$$

估计量 $\hat{f}_{\mathrm{HIST}}(x_0)$是以 x_0 为中心、箱子宽度为 $2h$ 的直方图估计值,因为它等于位于 x_0-h 与 x_0+h 之间的样本部分值被箱子宽度 $2h$ 去除。当 $\hat{f}_{\mathrm{HIST}}$在 x 范围内等分空间值 x 处计算时,即每一个为 $2h$ 单位,就得到直方图。

估计量 $\hat{f}_{\mathrm{HIST}}(x_0)$利用等权方式给出 $x_0\pm h$ 内所有观测值,即把式(9.1)重写为:

$$
\hat{f}_{\mathrm{HIST}}(x_0)=\frac{1}{Nh}\sum_{i=1}^{N}\frac{1}{2}\times\mathbf{1}\left(\left|\frac{x_i-x_0}{h}\right|<1\right) \tag{9.2}
$$

即使基本密度是连续的,这导致密度估计成为阶梯函数。比较光滑的估计值可通过利用加权函数来获得,而不是此处选用的指示函数。

9.3.2 核密度估计量

核密度估计量(**kernel density estimator**)是由罗森布拉特(Rosenblatt, 1956)引进的,通过利用可选择的加权函数对直方图估计(9.2)进行推广,因此:

$$
\hat{f}(x_0)=\frac{1}{Nh}\sum_{i=1}^{N}K\left(\frac{x_i-x_0}{h}\right) \tag{9.3}
$$

加权函数 $K(\cdot)$称为核函数(**kernel function**),并且满足下一节给出的约束。参数 h 是一个光滑参数,称为带宽(**bandwidth**),而 2 倍 h 是窗宽[2](**window width**)。与

[1] 又称为示性函数。——译者注
[2] 又称为窗孔宽度。——译者注

用于建立直方图的 x 值范围相比，密度通过在更广泛的 x_0 值处估计 $\hat{f}(x_0)$ 而得到；通常，计算是在样本值 $x_1, \cdots, x_N$ 处进行。这同样有助于提供比直方图更光滑的密度估计。

9.3.3 *核函数*

核函数 $K(\cdot)$ 是一个连续函数、关于零点对称，同时其积分为 1 且满足附加有界性条件。沿着李明宰（M. J. Lee，1996）的线索，我们假定核满足下述条件：

(i) $K(z)$ 关于 0 是对称的且连续的。

(ii) $\int K(z)dz = 1$，$\int zK(z)dz = 0$，并且 $\int |K(z)|dz < \infty$。

(iii) 或者(a)对于某个 z_0，当 $|z| \geq z_0$，则 $K(z) = 0$；或者(b)当 $|z| \to \infty$ 时，则 $K(z) \to 0$。

(iv) $\int z^2 K(z)dz = \kappa$，其中，κ 表示常值。

在实际应用中，核函数满足条件(iiia)而非较弱条件(iiib)时，核函数就会很好地起作用。然后，为了方便起见，将关注范围限制在 $[-1,1]$ 而不是 $[-z_0, z_0]$，这是正规化，而且通常将 $K(z)$ 限制在 $z \in [-1,1]$。

一些普遍使用的核函数已列于表 9.1 中。一致核利用相同权数作为直方图中的箱子宽度 $2h$，只是它所产生的直方图是利用一系列 x_0 点而不是利用固定箱子来计算的。高斯核满足(iiib)而不是(iiia)，因为它没有限制 $z \in [-1,1]$。p 阶核是指首次非零矩为第 p 阶的核。前 7 个核都二阶的，且满足条件(ii)中的第二个条件。最后 2 个核都是四阶核。如果 $f(x)$ 是多于二次可微的，那么这类较高阶核(**higher order kernel**)会使收敛速率增大（参见 9.3.10 节），尽管它们可取到负值。表 9.1 给出由式(9.11)定义且于 9.3.6 节运用的参数 δ，对某些核而言，有助于带宽的选择。

已知 $K(\cdot)$ 及 h，可相当简单地计算估计量。如果核估计量在距离 x_0 值 r 处计算，当核具有无界支持时，那么核估计量的计算至多需要 Nr 次运算。采用节省时间的计算方法是可行的；例如，参见哈德尔（Härdle，1990，第 35 页）。

9.3.4 *核密度例子*

对带宽 h 的一个重要选取已由图 9.2 阐述。

这里，利用图 9.1 的小时工资对数来阐述核的选择。[1]

图 9.4 显示利用各种不同核的效果。对于埃帕内尼科夫核、高斯核、四次核和一致核来说，由式(9.13)给出的西尔弗曼插值估计值所产生的带宽分别是 0.545、0.246、0.246 以及 0.214。甚至对于那些可产生直方图的一致核来说，所得到的核密度估计值都是非常相似的。带有核选择的密度估计变化远小于图 9.2 显示的带有不同带宽选择的变化。

---

[1] 原著中该段内容为 Here we illustrate the choice of kernel using …，应该将 using 及其后面的本段内容全部删掉。此处译者对该句做了更正。——译者注

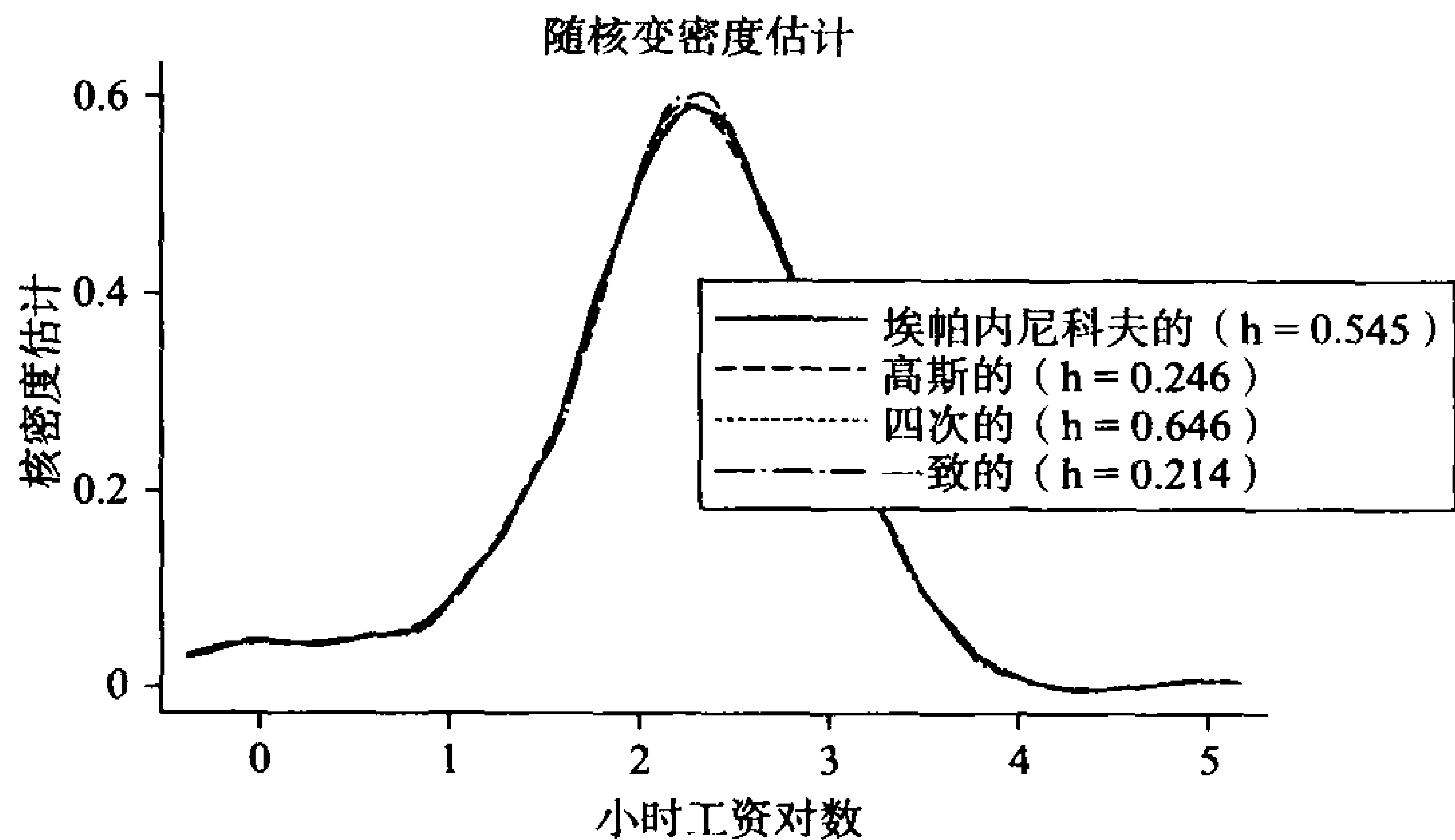


图 9.4 利用关于带宽的西尔弗曼插值估计，对四种不同的核，做出对数工资核密度估计。数据与图 9.1 的相同。

9.3.5 统计推断

假定数据是 iid 的，已知对 $K(\cdot)$与 h 的选取，我们来阐述核密度估计量 $\hat{f}(x)$ 的分布。该估计值 $\hat{f}(x)$是有偏的。如果当 $N\to\infty$时，带宽 $h\to0$，那么这个偏倚就渐近地趋于 0，因此，$\hat{f}(x)$是一致的。可是，偏倚项在 $\hat{f}(x)$的渐近正态分布中并不一定会消失，从而使统计推论变得复杂。

均值与方差

假定 $f(x)$ 的二阶导数存在且是有界的，同时核满足$\int zK(z)dz=0$，如同 9.3.3 节性质(ii) 所假定的，就可获得 9.8. 节中 $\hat{f}(x_0)$ 的均值与方差。

核密度估计量是有偏的，其偏倚项(**bias term**)$b(x_0)$ 依赖于带宽、真实密度曲率，并且依据

$$b(x_0)=\mathrm{E}[\hat{f}(x_0)]-f(x_0)=\frac{1}{2}h^2f''(x_0)\int z^2K(z)dz \tag{9.4}$$

所使用的核。核密度估计量偏倚量是 $O(h^2)$的，其中，我们使用了数量阶的记号，即如果 $a(h)/h^k$ 是有限的，那么函数 $a(h)$是 $O(h^k)$的。假如当 $N\to\infty$时有 $h\to0$，偏倚就将消失。

假定 $h\to0$ 且 $N\to\infty$，核密度估计量的方差(**variance**)是：

$$\mathrm{V}[\hat{f}(x_0)]=\frac{1}{Nh}f(x_0)\int K(z)^2dz+o\left(\frac{1}{Nh}\right) \tag{9.5}$$

其中，函数 $a(h)$表示 $o(h^k)$，当 $a(h)/h^k\to0$ 时。该方差依赖于样本量、真实密度以及核。当 $Nh\to\infty$，方差消失，这就要求当 $h\to0$ 时，方差必须以比 $N\to\infty$更慢的速率消失。

一致性

核估计量是逐点一致的(**pointwise consistent**)，也就是说，在特定点 $x=x_0$ 处是一致的，如果偏倚消失且方差也消失。这正是当 $h\to0$ 且 $Nh\to\infty$时的情况。

对于 $f(x)$在 x 处所有值的估计来说,较强的一致收敛(**uniform convergence**)条件,即 $\sup_{x_0}|\hat{f}(x_0)-f(x_0)|\xrightarrow{p}0$,可以证明,当 $Nh/\ln N\to\infty$时,会出现此情况。这要求 h 比逐点收敛的情况要大一些。

渐近正态性

前面结果表明,$\hat{f}(x_0)$渐近地具有均值 $f(x_0)+b(x_0)$和方差 $(Nh)^{-1}f(x_0)\times\int K(z)^2dz$。由此可得,应用中心极限定理,则核密度估计量具有极限分布(**limit distribution**):

$$\sqrt{Nh}(\hat{f}(x_0)-f(x_0)-b(x_0))\xrightarrow{d}\mathcal{N}\left[0,f(x_0)\int K(z)^2dz\right] \tag{9.6}$$

所应用的中心极限定理是非标准的,且需要条件(iv);例如,参见李明宰(Lee, 1996, 第 139 页)或帕甘和乌拉(Pagan and Ullah, 1999, 第 40 页)。

重要的是注意到,由式(9.4)定义的偏倚项 $b(x_0)$存在。就带宽的一般选取而言,这一项并不会消失,却使置信区间的计算变得复杂(将在 9.3.7 节阐述)。

9.3.6 *带宽选取*

选取带宽 h 的重要性远大于选取核函数 $K(\cdot)$的重要性。为减少偏倚而令 h 小一些与为确保光滑而令 h 大一些之间存在一种权衡的矛盾关系。因此,一种常规测量是均方误差(**mean-squared error, MSE**),即偏倚平方与方差之和。

由式(9.4)知,偏倚是 $O(h^2)$的,而由式(9.5)知,方差是 $O((Nh)^{-1})$的。从直观上看,通过选取 h 以使 MSE 极小化,因此,偏倚平方与方差是同阶的,所以 $h^4=(Nh)^{-1}$,这其中蕴含最优带宽 $h=O(N^{-0.2})$ 与 $\sqrt{Nh}=O(N^{0.4})$。现在,我们给出更正式的研究,包含 h 的实用插值估计。

均值积分平方误差

核密度估计在 x_0 处的效果局部(**local**)测量是:

$$\text{MSE}[\hat{f}(x_0)]=\text{E}[(\hat{f}(x_0)-f(x_0))^2] \tag{9.7}$$

其中,期望是关于密度 $f(x)$的。由于 MSE 等于方差加平方偏倚,所以由式(9.4)与式(9.5)得到核密度估计的 MSE:

$$\text{MSE}[\hat{f}(x_0)]\simeq\frac{1}{Nh}f(x_0)\int K(z)^2dz+\left\{\frac{1}{2}h^2f''(x_0)\int z^2K(z)dz\right\}^2 \tag{9.8}$$

为了获得在所有 x_0 值处的效果全局(**global**)测量,我们通过定义平方积分误差(**integrated squared error, ISE**)

$$\text{ISE}(h)=\int(\hat{f}(x_0)-f(x_0))^2dx_0 \tag{9.9}$$

来开始,在离散情况下,对所有 x_0 的平方误差进行求和,这一点类似于连续情况。这可写成 h 的函数,以此强调对带宽的依赖性。然后,我们除了对密度 $f(x)$取 ISE 的期望值之外,要去掉 $\hat{f}(x_0)$对 x 值而不是 x_0 的依赖性。从而,得出均值平方积

分误差(**mean integrated squared error, MISE**):

$$\begin{aligned}\mathrm{MISE}(h) &= \mathrm{E}[\mathrm{ISE}(h)] \\ &= \mathrm{E}\left[\int(\hat{f}(x_0)-f(x_0))^2 dx_0\right] \\ &= \int \mathrm{E}[(\hat{f}(x_0)-f(x_0))^2]dx_0 \\ &= \int \mathrm{MSE}[\hat{f}(x_0)]dx_0\end{aligned}$$

其中,$\mathrm{MSE}[\hat{f}(x)]$已由式(9.8)定义。由前面的代数运算知,MISE 等于积分均方误差(**integrated mean-squared error, IMSE**)。

最优带宽

最优带宽是求 MISE 极小值。对 MISE(h)求关于 h 的导数,并令其导数为 0,得到最优带宽(**optimal bandwidth**):

$$h^* = \delta\left(\int f''(x_0)^2 dx_0\right)^{-0.2} N^{-0.2} \tag{9.10}$$

其中,δ 依赖于所用的核函数,这里:

$$\delta = \left[\frac{\int K(z)^2 dz}{\left(\int z^2 K(z)dz\right)^2}\right]^{0.2} \tag{9.11}$$

该结果归功于西尔弗曼(Silverman, 1986)。

正如同一致性所需要的,由于 $h^* = O(N^{-0.2})$,故当 $N\to\infty$ 且 $Nh^* = O(N^{0.8})\to\infty$时,$h^*\to\infty$。$\hat{f}(x_0)$的偏倚是 $O(h^{*2}) = O(N^{-0.4})$,当 $N\to\infty$时它会消失。对于直方图估计来说,可以证明,$h^* = O(N^{-0.2})$且 $\mathrm{MISE}(h^*) = O(N^{-2/3})$,低于核密度估计的 $\mathrm{MISE}(h^*) = O(N^{-4/5})$。

最优带宽依赖于密度的曲率,当 $f(x)$是高度可变时,h^*就较小。

最优核

最优带宽会随核而变化[参见式(9.10)与式(9.11)],可以证明,倘若各种不同的最优 h^* 用于不同的核,MISE(h^*)则随核的不同而变动很小(图 9.4 提供一种阐述)。可以证明,最优核(**optimal kernel**)是埃帕内尼科夫核,尽管它的优点显得很少。

选取带宽的重要性远大于选取核的重要性,同时由式(9.10)知,这会随不同核而变化。

带宽插入估计

带宽的插入估计(**plug-in estimate**)是 h 的一个简单公式,这里,h 依赖于样本量 N 以及样本标准差 s。

一个有益的起点是,假定数据服从正态分布。于是,$\int f''(x_0)^2 dx_0 = 3/(8\sqrt{\pi}\sigma^5) = 0.211\,6/\delta^5$,在此情况下,式(9.10) 专门化为:

$$h^* = 1.3643\delta N^{-0.2} s \tag{9.12}$$

其中，s 表示 x 的样本标准差，而就几种核而言，δ 已列于表 9.1 中。对于埃帕内尼科夫核来说，$h^* = 2.345N^{-0.2}s$，而对于高斯核来说，$h^* = 1.059N^{-0.2}s$。就正态核而言，会出现相当小的带宽，因为与大多数核不同，当 $|x_i - x_0| > h$ 时，正态核会对 x_i 给出某种权数。在实际应用中，人们利用西尔弗曼插值估计(**Silverman's plug-in estimate**)：

$$h^* = 1.3643\delta N^{-0.2} \min(s,\ iqr/1.349) \tag{9.13}$$

其中，iqr 表示四分位数间距。这使用 $iqr/1.349$ 作为 σ 的一种可选择估计，以此预防异常值(**outliers**)，而这会使 s 增大，从而导致 h 非常大。

在实际应用中，h 的这些插入估计会很好地发挥作用，尤其是对于对称单峰密度，即使 $f(x)$ 不是正态密度。不过，人们还应通过利用各种变形来加以验证，譬如 2 倍插入估计与半个插入估计。

例如，在图 9.2 和图 9.4 中，我们有 $177^{-0.2} = 0.3551$，$s = 0.8282$，$iqr/1.349 = 0.6459$，因此，由式(9.13)得到，$h^* = 0.3173\delta$。例如，对埃帕内尼科夫核而言，得到 $h^* = 0.545$，因为由表 9.1 知，$\delta = 1.7188$。

交叉验证(cross-validation)

由式(9.9)知，$\mathrm{ISE}(h) = \int \hat{f}(x_0)dx_0 - 2\int \hat{f}(x_0)f(x_0)dx_0 + \int f^2(x_0)dx_0$。第三项不依赖于 h。一种可供选择的数据驱动方法是，通过

$$\mathrm{CV}(h) = \frac{1}{N^2 h}\sum_i \sum_j K^{(2)}\left(\frac{x_i - x_j}{h}\right) - \frac{2}{N}\sum_{i=1}^{N} \hat{f}_{-i}(x_i) \tag{9.14}$$

来估计 $\mathrm{ISE}(h)$ 的前两项，其中，$K^{(2)}(u) = \int K(u-t)K(t)dt$ 表示 K 对自身的卷积，而 $\hat{f}_{-i}(x_i)$ 表示 $f(x_i)$ 去掉一个核估计量。参见李明宰(Lee，1996，第 137 页)或帕甘和乌拉(Pagan and Ullah，1999，第 51 页)的推导。交叉验证估计(**cross-validation estimate**) h_{CV} 是求 $\widehat{\mathrm{CV}}(h)$ 极小值。可以证明，当 $N \to \infty$，$h_{CV} \xrightarrow{p} h^*$ 时，其收敛速度非常慢。

从计算上看，求 h_{CV} 显得很麻烦，因为 $\widehat{\mathrm{ISE}}(h)$ 需要在 h 值的范围内加以计算。人们时常不必去交叉验证核密度估计，因为插值估计通常提供了一个良好的起点。

9.3.7 置信区间

在没有置信区间时，通常要阐述核密度估计，但构造关于 $f(x_0)$ 的逐点置信区间是可能的，逐点意指在特定的 x_0 值处进行计算。一种简单方法(procedure，又称为程序)是，在计算点 x_0 的一个很小的数譬如 10 上获得置信区间，这恰好是 x 范围上的分布，并将它与估计密度曲线一起绘制成图。

由结果(9.6)可得到 $f(x_0)$ 的 95%置信区间(**confidence interval**)：

$$f(x_0) \in \hat{f}(x_0) - b(x_0) \pm 1.96 \times \sqrt{\frac{1}{Nh}\hat{f}(x_0)\int K(z)^2 dz}$$

就大多数核而言,很容易通过解析方法得到$\int K(z)^2dz$。

由于存在偏倚项,情况变得很复杂,这不应在有限样本下被忽略掉,尽管在渐近形式上$b(x_0)\xrightarrow{p}0$。这是因为含有最优带宽$h^*=O(N^{-0.2})$,由式(9.6)给出的重新标度随机变量($\sqrt{Nh}(\hat{f}(x_0)-f(x_0))$)的偏倚并没有消失,由于$\sqrt{Nh^*}$乘以$O(h^*)=O(1)$,由式(9.4)与$f''(x_0)$的核估计能估计出该偏倚,可是在实际应用中,估计$f''(x_0)$就显得繁琐。相反,通常方法是减少计算置信区间的偏倚而不是$\hat{f}(x_0)$自身,这里要通过光滑不足来完成,即选取$h<h^*$,因而$h^*=o(N^{-0.2})$。另外一些方法包括,利用较高阶核,例如由表9.1给出的四阶核或自助法(参见11.6.5节)。

同样地,人们能对x的所有可能值计算$f(x)$的置信带。与每个x_0值的逐点置信区间相比,这些置信带要宽一些。

9.3.8 密度导数的估计

在一些情况下,需要对密度导数(**derivatives**)进行估计。例如,对式(9.4)给出的$\hat{f}(x_0)$的偏倚项进行估计时,要求估计$f''(x_0)$。

为了简单起见,我们阐述一阶导数的估计。有限差分方法是使用$\hat{f}'(x_0)=[\hat{f}(x_0+\Delta)-\hat{f}(x_0-\Delta)]/2\Delta$。而微分方法则求式(9.3)中$\hat{f}(x_0)$的一阶导数,得到$\hat{f}'(x_0)=-(Nh^2)^{-1}\sum_i K'((x_i-x_0)/h)$。

从直观上讲,较大带宽应该用于估计导数,这比$f(x_0)$更易变化。$\hat{f}^{(s)}(x_0)$的偏倚如前所述,只是方差收敛更慢一些,如果$f(x_0)$是p次可微的,那么会得到最优带宽$h^*=O(N^{-1/(2s+2p+1)})$。对于一阶导数核估计来说,我们需要$p\geqslant 3$。

9.3.9 多变量核密度估计

前面的讨论已经考察纯量x的核密度估计。对于k维随机变量$\mathbf{x}$的密度来说,多元核密度估计量(**multivariate kernel density estimator**)是:

$$\hat{f}(\mathbf{x}_0)=\frac{1}{Nh^k}\sum_{i=1}^{N}K\left(\frac{\mathbf{x}_i-\mathbf{x}_0}{h}\right)$$

其中,$K(\cdot)$现在表示k维核。通常$K(\cdot)$表示积核(**product kernel**),即一维核乘积。也可以使用多变量核,譬如多元正态密度或者与$K(\mathbf{z}'\mathbf{z})$成比例的球核。核$K(\cdot)$满足的性质类似于一维情况下得到的一些性质;参见李明宰(Lee, 1996,第125页)。

其解析结果与表达式均类似于前面一维的情况,只是$\hat{f}(\mathbf{x}_0)$的方差以速率$O(Nh^k)$下降,对于$k>1$来说,这比一维情况要慢一些。于是有:

$$\sqrt{Nh^k}(\hat{f}(\mathbf{x}_0)-f(\mathbf{x}_0)-b(\mathbf{x}_0))\xrightarrow{b}\mathcal{N}\left[0,f(\mathbf{x}_0)\int K(\mathbf{z})^2d\mathbf{z}\right]$$

最优带宽选择是$h=O(N^{-1/(k+4)})$,这比一维情况下$O(N^{-0.2})$的要大一些,并蕴含$\sqrt{Nh^k}=O(N^{2/(k+4)})$。插值方法与交叉验证方法能推广到多变量情况。就正态积

核而言，**x** 的第 j 个分量的斯科特插值(**Scott's plug-in**)估计是 $h_j = N^{-1/(k+4)} s_j$，其中，s_j 表示 x_j 的样本标准差。

对于多元变量核来说，更可能产生数据的稀疏性(**sparseness**)问题。当 **x** 具有较高维数时，由于 $\mathbf{x}_0$ 附近的少数几个观测值接收到大量的权数，故存在维数祸根。甚至这并不是一个问题，绘制二元核密度估计还需要三维曲线图，而这种曲线图很难被人们看懂与解释。

多元变量核密度估计的一种使用是，允许对条件密度进行估计。由于 $f(y|x)=f(x,y)/f(x)$，所以一种明显的估计量是，$\hat{f}(y|x)=\hat{f}(x,y)/\hat{f}(x)$，其中，$\hat{f}(x,y)$ 与 $\hat{f}(x)$分别是二变量与单变量的核密度估计值。

9.3.10 *较高阶核*

前面分析假定 $f(x)$是二次可微的，这是获得式(9.4)中偏倚项所必需的假设。若 $f(x)$是多于二次可微的，则利用较高阶核(参见 9.3.3 节中四阶核的例子)可减少偏倚阶数，从而使 h^* 较小且有较快的收敛速率。一般陈述是，如果 **x** 是 k 维的且 $f(\mathbf{x})$是 p 次可微的，同时使用第 p 阶核，那么 $f(\mathbf{x})$的核估计 $\hat{f}(\mathbf{x}_0)$具有最优收敛速率 $N^{-p/(2p+k)}$，其中，$h^* = O(N^{-1/(2p+k)})$。

9.3.11 *可选择的非参数密度估计*

核密度估计是一种标准的非参数估计。例如，帕甘和乌拉(Pagan and Ullah, 1999)曾阐述其他的密度估计。密度估计经常使用诸如最近邻方法，该方法更普遍用于非参数回归之中，并在 9.6 节加以简略阐述。

9.4 非参数局部回归

我们考察纯量因变量 y 对纯量回归元变量 x 的回归。该回归模型是：

$$\begin{aligned} y_i &= m(x_i)+\varepsilon_i, \quad i=1,\cdots,N \\ \varepsilon_i &\sim \text{iid}[0,\sigma_\varepsilon^2] \end{aligned} \tag{9.15}$$

其复杂性在于函数形式 $m(\cdot)$没有被设定，因此，非线性最二小乘法估计不可行。

本节提供一种简单利用局部加权平均(**local weighted average**)对非参数回归(**nonparametric regression**)做出的一般性研究。对核回归的专门研究，将在 9.5 节给出，而其他普遍使用的局部加权方法则在 9.6 节中加以阐述。

9.4.1 *局部加权平均*

假定对于回归元的单个值譬如 x_0，y 存在多重观测值，比如说 y 有 N_0 个观测值。于是，$m(x_0)$的一个明显而简单的估计量是，对 y 的这些 N_0 个值进行简单平均，我们将其记为 $\tilde{m}(x_0)$。由此可得，$\tilde{m}(x_0)\sim[m(x_0), N_0^{-1}\sigma_\varepsilon^2]$，由于它是 N_0 个观测值的平均，由式(9.15)可知，$\tilde{m}(x_0)$是 iid 的，且均值为 $m(x_0)$、方差为 σ_ε^2。

估计量 $\tilde{m}(x_0)$是无偏的，但不一定是一致的。一致性要求，当 $N\to\infty$时有

$N_0 \to \infty$，因此 $V[\tilde{m}(x_0)] \to 0$。就离散回归元而言，在有限样本下，该估计量非常繁琐，因为 N_0 可以是很小的。甚至更糟的是，就连续回归元而言，对 x_i 取特殊值 x_0 的情况来说，甚至当 $N \to \infty$ 时，仅仅存在一个观测值。

除 x 确实等于 x_0 以外，当 x 接近于 x_0 时，数据的稀疏性问题可通过对 y 的观测值进行平均加以克服。我们注意到，估计量 $\tilde{m}(x_0)$ 能表示成因变量的加权平均，即 $\tilde{m}(x_0)=\sum_i w_{i0} y_i$，其中，当 $x_i=x_0$ 时，权数 w_{i0} 等于 $1/N_0$，而当 $x_i \neq x_0$ 时，权数等于 0。因此，权数既随计算点 x_0 变化，又随回归元的样本值变化。

更一般地讲，我们考察局部加权平均估计量（**local weighted average estimator**）：

$$\hat{m}(x_0) = \sum_{i=1}^{N} w_{i0,h} y_i \tag{9.16}$$

其中，权数：

$$w_{i0,h} = w(x_i, x_0, h)$$

之和为 1，所以 $\sum_i w_{i0,h}=1$。对权数设定为，当 x_i 越接近 x_0 时，其值越大。

另一个参数 h 为窗口宽度参数[1]（**window width parameter**）的一般性符号。对它的定义是，愈小的 h 值导致窗口愈小，并且对 x_i 接近于 x_0 的那些观测值，则设定更大权数。在特定核回归例子中，h 表示带宽。9.6 节给出的其他一些方法具有可供选择的光滑参数（**smoothing parameters**），光滑参数起着类似于 h 的作用。当 h 变小时，$\hat{m}(x_0)$ 变得稍许有些偏倚，因为仅有接近 x_0 的观测值才能被使用，却更容易变化，其原因在于使用很少的观测值。

线性回归模型的普通最小二乘预测式是 y_i 的加权平均，因为经过一些代数运算，可以得到：

$$\hat{m}_{\mathrm{OLS}}(x_0) = \sum_{i=1}^{N} \left\{ \frac{1}{N} + \frac{(x_0-\bar{x})(x_i-\bar{x})}{\sum_j (x_j-\bar{x})^2} \right\}$$

可是，例如当 $x_i > x_0 > \bar{x}$ 时，普通最小二乘权数实际上会随着 x_0 与 x_i 之间距离增大而递增。相反，局部回归（**local regression**）则使用随 $|x_i - x_0|$ 而递减的权数。

9.4.2 *K 最近邻例子*

我们考察一个简单例子，对应于最接近于 x 且小于 x_0 的 $(k-1)/2$ 个观测值与最接近于 x 且大于 x_0 的 $(k-1)/2$ 个观测值的 y 值进行未加权平均。

通过增大 x 值的方式，对观测值排序。然后，在 $x_0=x_i$ 处加以计算，得到：

$$\hat{m}_k(x_i) = \frac{1}{k}(y_{i-(k-1)/2} + \cdots + y_{i+(k-1)/2})$$

其中，为了简单起见，k 表示奇数，并通过一些联系进行潜在修改，同时忽略接近于端点 x_1 或 x_N 的 x_0 之值。该估计量可被表述成式(9.16)的一种特殊情况，其权数为：

$$w_{i0,k} = \frac{1}{k} \times \mathbf{1}\left(|i-0| < \frac{k-1}{2}\right), \quad x_1 < x_2 < \cdots < x_0 < \cdots < x_N$$

---

[1] 又称为窗宽参数。——译者注

这个估计量有许多称谓。我们把它称为(对称的)k 最近邻估计量(**k-nearest neighbors estimator**，记为 k - NN),它已在 9.6.1 节定义。它也可以是一种以 x_0 为中心、长度为 k 的局部进行平均(**local running average**),或求均值(**running mean**),或求移动平均(**moving average**)。例如,绘制时间序列 y 与时间 x 的曲线图。参数 k 起着 9.4.1 节中窗口宽度 h 的作用,小 k 对应于小 h。

举一个例子阐述,考察源自模型:

$$
\begin{aligned}
& y_i = 150 + 6.5x_i - 0.15x_i^2 + 0.001x_i^3 + \varepsilon_i, \quad i = 1, \cdots, 100 \qquad (9.17) \\
& x_i = i \\
& \varepsilon_i \sim \mathcal{N}[0, 25^2]
\end{aligned}
$$

的生成数据。y 的均值关于 x 是三次的,x 取值为 1,2,…,100,其转向点在 $x=20$ 与 $x=80$ 处。为此,要增加服从正态分布的误差项,其标准差为 25。

图 9.5 绘制出满足 $k=5$ 及 25 的对称 k - NN 估计量。这两种移动平均都建议三次关系。第二个比第一个更光滑一些,但仍是相当凸凹不平,尽管样本的 1/4 已用于形成平均值。普通最小二乘回归线也画在此图中。

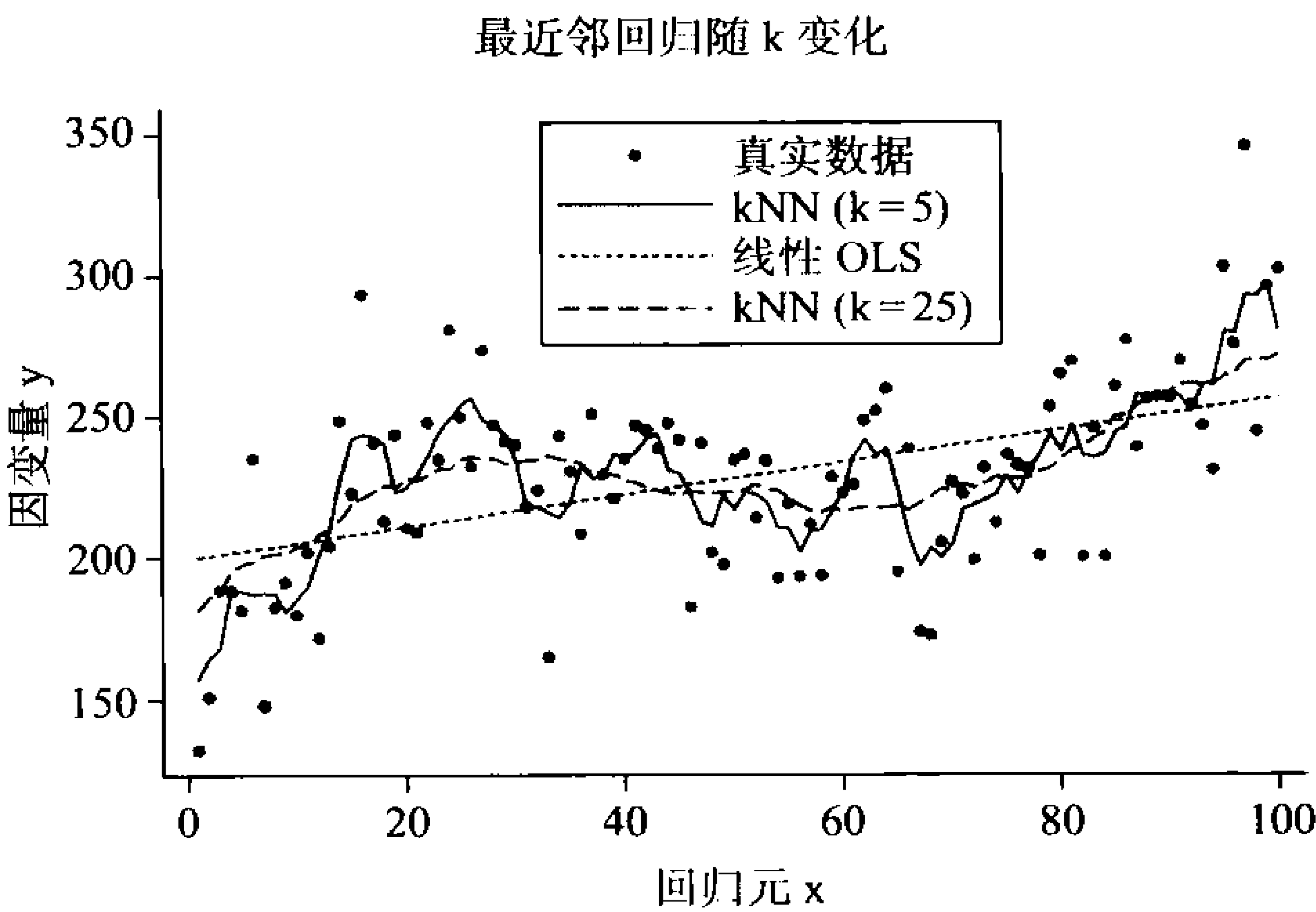


图 9.5 对 k 的两种不同选取,画出 k 最近邻回归曲线与 OLS 回归线。数据是由三次多项式模型生成的。

当 $k=25$ 而不是 $k=5$ 时,$\hat{m}_k(x)$在端点外的斜率更为平坦一些。这就阐明了在端点处估计 $m(x)$时会出现边界问题(**boundary problem**)。例如,对于最小回归元值 x_1 来说,不存在包括 x 的下方观测值,而平均则变成一个单侧平均 $\hat{m}_k(x_1) = (y_1 + \cdots + y_{1+(k-1)/2})/[(k+1)/2]$。就这些数据而言,由于 $m_k(x)$在该区域内关于 x 是递增的,所以导致 $\hat{m}_k(x_1)$被过高估计,而过分夸大关于 k 是递增的。相反,这类边界问题可通过利用 9.6.2 节给出的方法加以减少。

9.4.3 *洛斯回归例子*

一旦将可供选择的权数用到那些建立对称化的 k - NN 估计量上,就能得出更好的 $m(x)$估计值。

一个例子是 9.6.2 节定义的洛斯估计量(Lowess estimator)。它提供了 $m(x)$ 的一个较光滑估计,因为它使用核权数而不是指示函数,类似于核密度估计,比实施直方图更为光滑些。它还具有较小的偏倚(参见 9.6.2 节),这特别有利于在端

点处估计 $m(x)$。

对于由式(9.17)生成的数据来说，图 9.6 绘制出满足 $k=25$ 的洛斯估计值。这种局部回归估计相当于接近真实三次条件均值函数，该三次条件均值函数也被绘制出来。一旦将图 9.6 与 $k=25$ 的对称化 k - NN 的图 9.5 进行比较，可以看到，洛斯回归产生了更光滑的回归函数估计以及在边界处更准确的估计。

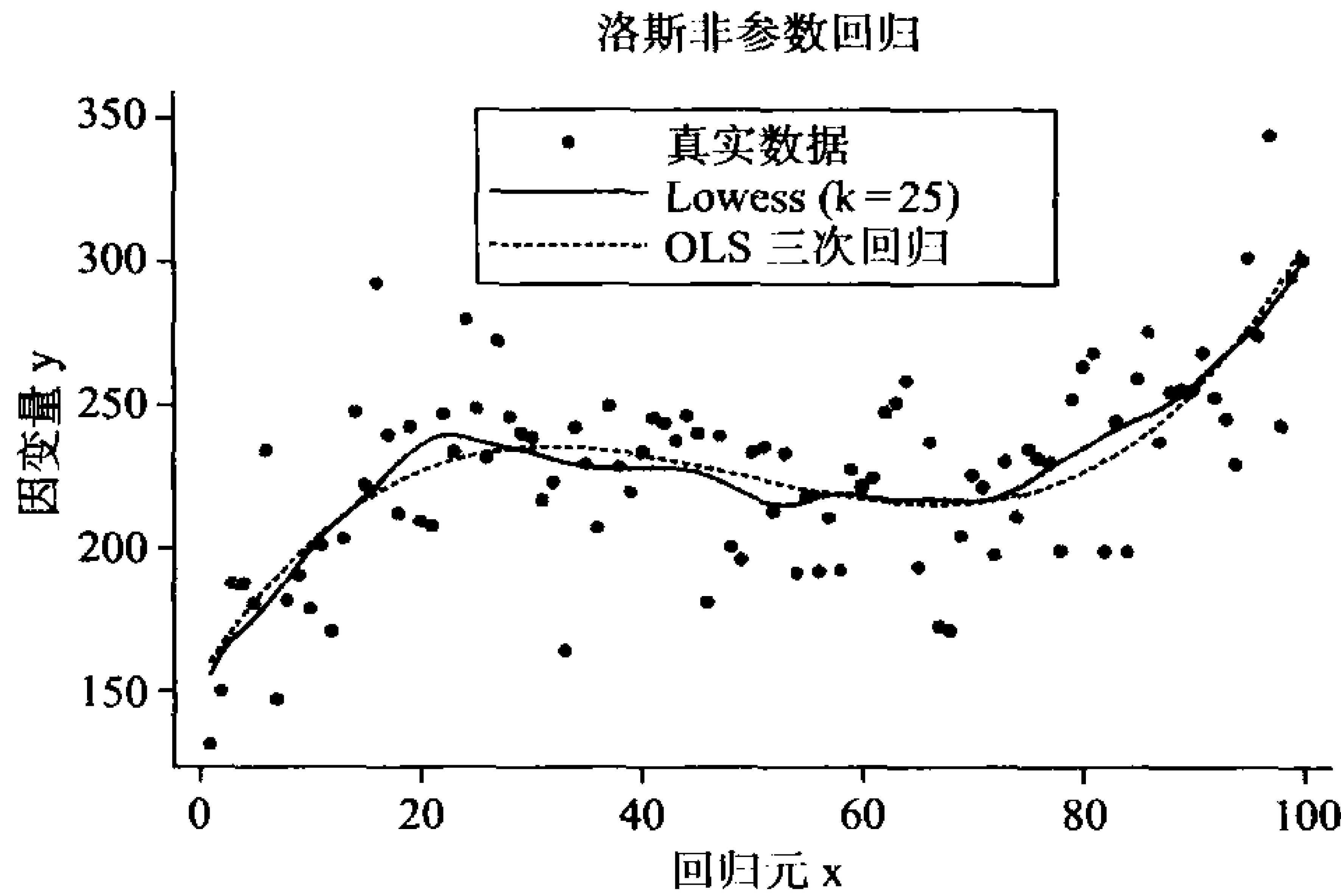


图 9.6 利用洛斯曲线及三次回归曲线的非参数回归曲线。数据生成过程与图 9.5 的一样。

9.4.4 *统计推断*

当误差项服从正态分布，并且分析是以 $x_1,\cdots,x_N$ 为条件时，很容易获得式(9.16)中 $\hat{m}(x_0)$ 的准确小样本分布。一旦把 $y_i=m(x_i)+\varepsilon_i$ 代入 $\hat{m}(x_0)$ 的定义中，就会直接得到：

$$\hat{m}(x_0)-\sum_{i=1}^{N}w_{i0,h}m(x_i)=\sum_{i=1}^{N}w_{i0,h}\varepsilon_i$$

对于固定回归元且如果 ε_i 服从 iid $\mathcal{N}[0,\sigma_\varepsilon^2]$，这蕴含：

$$\hat{m}(x_0)\sim\mathcal{N}\left[\sum_{i=1}^{N}w_{i0,h}m(x_i),\ \sigma_\varepsilon^2\sum_{i=1}^{N}w_{i0,h}^2\right] \tag{9.18}$$

注意到，通常 $\hat{m}(x_0)$ 是有偏的，且其分布不一定以 $m(x_0)$ 为中心。

如果拥有随机回归元和非正态误差，我们以 $x_1,\cdots,x_N$ 为条件，并应用 U 统计量的中心极限定理，U 统计量是适合双重求和的[例如，参见帕甘和乌拉(Pagan and Ullah,1999,第 359 页)]。于是，对于 ε_i 服从 iid $[0,\sigma_\varepsilon^2]$，有：

$$c(N)\sum_{i=1}^{N}w_{i0,h}\varepsilon_i\xrightarrow{d}\mathcal{N}\left[0,\ \sigma_\varepsilon^2\lim c(N)^2\sum_{i=1}^{N}w_{i0,h}^2\right] \tag{9.19}$$

其中，$c(N)$ 表示样本量的函数，满足 $Oc(N)<N^{1/2}$，$Oc(N)$ 随局部估计量而变化。例如，关于核回归，$c(N)=\sqrt{Nh}$，而对于具有最优带宽的核回归，$c(N)=N^{0.4}$。从而有[1]：

$$c(N)(\hat{m}(x_0)-m(x_0)+b(x_0))\xrightarrow{d}\mathcal{N}\left[0,\ \sigma_\varepsilon^2\lim c(N)^2\sum_{i=1}^{N}w_{i0,h}^2\right] \tag{9.20}$$

[1] 该式中 $b(x_0)$ 前面原文为"−"，但应为"+"。——译者注

其中，$b(x_0)=m(x_0)-\sum_i w_{i0,h}m(x_i)$。注意到，由式(9.20)得出 $\hat{m}(x_0)$ 的渐近分布(9.18)。

很明显，$\hat{m}(x_0)$ 的分布，即简单加权平均，能在可供选择的分布假设下获得。例如，就异方差误差(**heteroskedastic errors**)而言，式(9.19)与式(9.20)中的方差可用 $\lim c(N)^2\sum_i\sigma_{\varepsilon,i}^2 w_{i0,h}^2$ 来代替，这能通过用平方残差 $(y_i-\hat{m}(x_i))^2$ 代替 $\sigma_{\varepsilon,i}^2$ 而得到一致估计。一种可供选择的方式是，人们能够运用自助法(参见 11.6.5 节)。

9.4.5 选取带宽

本章我们沿着非参数技术路线，即如果 $\hat{\theta}=\theta_0+O_p(N^{-r})$，那么 θ_0 的估计量具有收敛速率(**convergence rate**) N^{-r}，因此，$N^r(\hat{\theta}-\theta_0)=O_p(1)$，而且原则上 $N^r(\hat{\theta}-\theta_0)$ 具有极限分布。注意，尤其是被广泛称为 $\sqrt{N}$ 一致估计量的那种估计量以速率 $N^{-1/2}$ 收敛。与该估计量相比，非参数估计量典型地表现出较慢的收敛速率，即 $r<1/2$，因为需要用小的带宽 h 消除偏倚，从而可用少于 N 个观测值估计 $\hat{m}(x_0)$。

举一个例子，考察 9.4.2 节的 k - NN 例子。假定 $k=N^{4/5}$，因而当 $N=1\ 000$ 时，$k=251$。于是，此估计量是一致的，由于移动平均使用了样本的 $N^{4/5}/N=N^{-1/5}$ 个观测值，所以当 $N\rightarrow\infty$ 时，在 x_0 附近失效。一旦利用式(9.18)，移动平均估计量的方差是 $\sigma_\varepsilon^2\sum_i w_{i0,k}^2=\sigma_\varepsilon^2\times k\times(1/k)^2=\sigma_\varepsilon^2\times 1/k=\sigma_\varepsilon^2 N^{-4/5}$，所以式(9.19)中的 $c(N)=\sqrt{k}=\sqrt{N^{4/5}}=N^{0.4}$，小于 $N^{1/2}$。倘若 $k<O(N)$，则 k 的其他值同样能够保证一致性。

更一般地讲，一系列带宽参数的值都会消除渐近偏倚，但较小带宽会增加可变性。在这方面文献中，对这种权衡可通过对均方误差求极小值、方差之和以及偏倚平方来解释。

斯通(Stone，1980)已经证明，当 $\mathbf{x}$ 是 k 维的，同时 $m(\mathbf{x})$ 是 p 次可微的，则 $m(\mathbf{x})$ 的第 s 阶导数非参数估计量的可能最快收敛速率是 N^{-r}，其中，$r=(p-s)/(2p+k)$。该速率随着导数阶数增大且 $\mathbf{x}$ 维数增加而递减。这个速率会随着假定 $m(\mathbf{x})$ 可微次数的增大而递增，如果 $m(\mathbf{x})$ 具有趋于无穷次阶数的导数，那么它趋于 $N^{1/2}$。对于 $m(\mathbf{x})$ 的纯量回归估计来说，一种习惯做法是，假定 $m''(\mathbf{x})$ 存在，在此情况下，$r=2/5$ 且最快收敛速率为 $N^{-0.4}$。

9.5 核回归

核回归是利用核权数的一种加权平均估计量。一些问题譬如对核密度估计所阐述的偏倚及带宽选取同样与核回归有关。不过，与回归情况相比，对带宽选取的关注显然不足。同样地，尽管我们为了教学原因而阐述核回归，但在实际应用时经常使用非核局部回归估计量[1](参见 9.6 节)。

---

[1] 原著此处为“核局部回归估计量”，但应该为“非核局部回归估计量”，这里译者已做了相应改动。——译者注

9.5.1 核回归估计

核回归的目的是要估计由式(9.15)定义的模型 $y=m(x)+\varepsilon$ 里面的回归函数 $m(x)$。由 9.4.1 节知，$m(x_0)$的一个明显估计量是，对应于接近 x_0 的 x_i 的因变量的样本值 y_i 的平均值。在此之上的一种变形是，求出关于距离 x_0 为 h 之内的 x_i 所有观测值的 y_i 的平均值。这正式地表述成：

$$\hat{m}(x_0)\equiv\frac{\sum_{i=1}^{N}\mathbf{1}\left(\left|\frac{x_i-x_0}{h}\right|<1\right)y_i}{\sum_{i=1}^{N}\mathbf{1}\left(\left|\frac{x_i-x_0}{h}\right|<1\right)}$$

其中，如前所述，当事件 A 发生，$\mathbf{1}(A)=1$，否则 $\mathbf{1}(A)=0$。分子对 y 值求和，而分母则给出求和时 y 值的个数。

这种表达式对接近 x_0 的所有观测值给出了相等权数，但一种可能更为受欢迎的方法是，在 x_0 处给出最大权数而远离 x_0 的权数则递减。因而，更一般地讲，我们考察 9.3.2 节引入的核加权函数 $K(\cdot)$。这就得到核回归估计量(**kernel regression estimator**)：

$$\hat{m}(x_0)\equiv\frac{\frac{1}{Nh}\sum_{i=1}^{N}K\left(\frac{x_i-x_0}{h}\right)y_i}{\frac{1}{Nh}\sum_{i=1}^{N}K\left(\frac{x_i-x_0}{h}\right)}\tag{9.21}$$

几种通用核回归，比如一致核回归、高斯核回归、埃帕内尼科夫核回归以及二次核回归，均列在表 9.1 中。

常数 h 称为带宽(**bandwidth**)，并将 $2h$ 称为窗口宽度(**window width**)。带宽所起的作用与 9.4.2 节 k - NN 例子中 k 的作用一样。

估计量(9.21)是由纳达雷娅(Nadaraya, 1964)与沃森(Watson, 1964)提出的，他们给出了一种可供选择的推导。条件均值 $m(x)=\int yf(y|x)dy=\int y[f(y,x)/f(x)]dy$，这能由 $\hat{m}(x)=\int y[\hat{f}(y,x)/\hat{f}(x)]dy$ 估计出来，其中，$\hat{f}(y,x)$与$\hat{f}(x)$是两变量核密度估计量与单变量核密度估计量。可以证明，这等于式(9.21)中的估计量。统计文献还考虑固定设计(**fixed design**)或固定回归元情况下的核回归，其中，$f(x)$是已知的且不需要估计，不过，我们只考察由观测数据引起的随机回归元(**stochastic regression**)情况。

核回归估计量是加权平均式(9.16)的一种特殊情况，其权数为：

$$w_{i0,h}=\frac{\frac{1}{Nh}K\left(\frac{x_i-x_0}{h}\right)}{\frac{1}{Nh}\sum_{i=1}^{N}K\left(\frac{x_i-x_0}{h}\right)}\tag{9.22}$$

由构建方法知，对 i 求和为 1。尽管这与 9.4 节的一般结果有关，但我们仍给出更详细的分析。

9.5.2 统计推断

一旦假定数据是 iid 的，已知对 $K(\cdot)$ 及 h 的选取，我们阐述核回归估计量 $\hat{m}(x)$的分布。我们以隐性方式假定回归元都是连续的。对离散回归元来说，$\hat{m}(x_0)$仍将在 $m(x_0)$ 处失效，同时 $\hat{m}(x_0)$ 极限形式和 $m(x_0)$都是阶梯函数。

一致性

对于条件均值函数 $m(x_0)$来说，$\hat{m}(x_0)$的一致性(**consistency**)要求 $h\to 0$，因此，大的权数只给予非常接近 x_0 的 x_i。可是，我们需要接近 x_0 的众多 x_i，因此，许多观测值用于建立加权平均。正式地讲，如果当 $N\to\infty$时，$h\to 0$ 且 $Nh\to\infty$，那么 $\hat{m}(x_0)\xrightarrow{p}m(x_0)$。

偏倚

假定 $m(x)$是二次可微的，则核回归估计量是有偏的，其偏倚数量是 $O(h^2)$的，偏倚项是：

$$b(x_0)=h^2\left(m'(x_0)\frac{f'(x_0)}{f(x_0)}+\frac{1}{2}m''(x_0)\right)\int z^2K(z)dz \tag{9.23}$$

(参见 9.8.2 节)。就核密度估计而言，其偏倚会随着所使用的核函数而变化。更为重要的是，偏倚依赖于回归函数 $m(x_0)$的斜率与曲率，还有回归元的密度 $f(x_0)$的斜率，而就密度估计而言，偏倚仅依赖于 $f(x_0)$的二阶导数。偏倚在端点处表现得特别大，正如 9.4.2 节所阐明的。

通过利用 9.3.3 节定义的较高阶核，以及边界修改譬如特定边界核，就能减少偏倚。局部多项式回归以及譬如洛斯(参见 9.6.2 节)修改都颇为引人注目，即去掉式(9.23)中依赖于 $m'(x_0)$的项，从而在边界上表现得很好。

渐近正态性

在 9.8.2 节，已经证明，对于具有密度 $f(x_i)$的 iid 的 x_i 来说，核回归估计量具有极限分布(**limit distribution**)：

$$\sqrt{Nh}(\hat{m}(x_0)-m(x_0)-b(x_0))\xrightarrow{d}\mathcal{N}\left[0,\frac{\sigma_\varepsilon^2}{f(x_0)}\int K(z)^2dz\right] \tag{9.24}$$

对小的 $f(x_0)$来说，式(9.24)中的方差(**variance**)却较大，因此，如同人们所料，在 x 稀少的区域上，$\hat{m}(x_0)$的方差却较大。

9.5.3 带宽选择

一旦把关于 $x_i\neq x_0$ 的那些 y_i 值并入加权平均之中，就会传入偏倚，因为对于 $x_i\neq x_0$，$\mathrm{E}[y_i|x_i]=m(x_i)\neq m(x_0)$。不过，由于我们利用更多数据加以平均，故利用这些额外点会减少估计量的方差。最优带宽是在增大偏倚与减少方差之间做出一种权衡，这用到了平方误差损失。与核密度估计不同，插入值方法并不切合实际，而交叉验证则应用更为广泛。

为了简单起见，大多数研究都关注为 x_0 所有值选取一个带宽。一些方法具有可变的带宽，像著名的 k - NN 与洛斯方法，这都在 9.6 节给出。

积分均方误差

$\hat{m}(\cdot)$在 x_0 点的局部效果是，用均方误差（**mean-squared error**）来测量，它由：

$$\mathrm{MSE}[\hat{m}(x_0)]=\mathrm{E}[(\hat{m}(x_0)-m(x_0))^2]$$

给出，其中，期望消除了 $\hat{m}(x_0)$对 x 的依赖性。由于 MSE 等于方差加平方偏倚，故能够利用式(9.23)与式(9.24)来获得 MSE。

与 9.3.6 节相类似，平方积分误差(**integrated square error**)是：

$$\mathrm{ISE}(h)=\int(\hat{m}(x_0)-m(x_0))^2 f(x_0)dx_0$$

其中，$f(x)$表示回归元 x 的密度，而均值平方积分误差(**mean integrated square error**)或等价形式的积分均方误差是：

$$\mathrm{MISE}(h)=\int\mathrm{MSE}[\hat{m}(x_0)]f(x_0)dx_0$$

最优带宽

最优带宽 h^* 是求 MISE(h)极小值。这就得到，$h^*=O(N^{-0.2})$，因为由式(9.23)知，偏倚是 $O(h^2)$；而由式(9.24)知，方差是 $O((Nh)^{-1})$，因为在用$\sqrt{Nh}$标度 $\hat{m}(x_0)$之后，可获得 $O(1)$方差；为使偏倚平方与方差成为同阶形式，则$(h^2)^2=(Nh)^{-1}$或者 $h=N^{-0.2}$。于是，核估计以速率 $(Nh^*)^{-1/2}=N^{-0.4}$收敛到 $m(x_0)$，而不是以参数分析中的通常速率 $N^{-0.5}$收敛。

插入带宽估计

如果利用与 9.3.5 节中关于核密度估计量的那种方法相类似的微分方法，人们就能获得求 MISE(h)极小值的 h^*准确表达式。于是，h^*依赖于式(9.23)与式(9.24)中的偏倚及方差。

插入方法(**plug-in approach**)是利用这些未知的估计值来计算 h^*。不过，例如，对 $m''(x)$进行估计时需要非参数方法，该非参数方法同样需要初始带宽选取，但 h^*还是依赖于一些未知量譬如 $m''(x)$。出现这些复杂情况时，人们就应该慎用插入估计。一种更为通行的方法是运用交叉验证，该方法将在下面阐述。

而且，可以证明，当使用埃帕内尼科夫核[参见哈德尔(Härdle，1990，第 186 页)；或哈德尔和林顿(Härdle and Linton，1994，第 2 321 页)]时，MISE(h^*)就被极小化了，尽管如同核回归一样，MISE(h^*)对其他核来说并不是较大的。一个关键性问题是确定 h^*，这将会随核与数据而变化。

交叉验证

通过去掉一个交叉验证(**cross-validation**)方法，可获得最优 h 的经验估计。该方法是选取 $\hat{h}^*$，以使下式极小化：

$$\mathrm{CV}(h)=\sum_{i=1}^{N}(y_i-\hat{m}_{-i}(x_i))^2\pi(x_i) \tag{9.25}$$

其中，$\pi(x_i)$表示加权函数(下面将讨论)，而：

$$\hat{m}_{-i}(x_i) = \sum_{j\neq i} w_{ji,h} y_j \Big/ \sum_{j\neq i} w_{ji,h} \tag{9.26}$$

表示通过核公式(9.21)或更一般地通过加权方法(9.16)式，连同去掉 y_i 的调整而获得的 $m(x_i)$ 的去掉一个估计值(**leave-one-out estimate**)。

正如交叉验证第一次出现的那样，交叉验证不是一种密集计算。可以证明：

$$y_i - \hat{m}_{-i}(x_i) = \frac{y_i - \hat{m}(x_i)}{1 - [w_{ii,h}/\sum_j w_{ji,h}]} \tag{9.27}$$

因此，对于每一个 h 值，交叉验证仅仅需要加权平均 $\hat{m}(x_i)$ 的一种计算，$i=1,\cdots,N$。

为了潜在降低端点权数，要引入权数 $\pi(x_i)$，否则可能是一个严重的问题，如同9.4.2节所阐述的，原因在于端点处局部加权估计具有极高的偏倚。例如，x_i 外面的第5%百分位数到第95%百分位数观测值并没有用于计算 CV(h)，在此情况下，对这些观测值来说，$\pi(x_i)=0$，而其他情况下，$\pi(x_i)=1$。运用交叉验证术语是因为它证实了利用数据集中所有其他观测值去预测第 i 个观测值的能力。第 i 个观测值被忽略掉，是因为如果在预测中额外地使用它，当 $\hat{m}_h(x_i)=y_i(i=1,\cdots,N)$ 时，CV(h)自然被极小化。CV(h)也称为估计预测误差(**estimated prediction error**)。

哈德尔和马伦(Härdle and Marron, 1985)已经证明，对 CV(h)求极小值，渐近地等价于对 ISE(h)的修正值与 MISE(h)求极小化。这种修正包括被积函数中的加权函数 $\pi(x_0)$，以及平均平方误差(**averaged squared error**, ASE) $N^{-1}\sum_i(\hat{m}(x_i)-m(x_i))^2\pi(x_i)$，这是对 ISE($h$)的一种离散样本近似。不过，测量 CV($h$)以低速率 $O(N^{-0.1})$ 收敛，因此，CV(h)在有限样本中表现出相当可变的。

广义交叉验证

对去掉一个交叉验证来说，一种可供选择的方法是使用类似于 CV(h)的测量，只是更简单地使用 $\hat{m}(x_i)$ 而不是 $\hat{m}_{-i}(x_i)$，然后增加模型复杂性惩罚，该惩罚会随着带宽 h 减小而增大。从而得出：

$$\mathrm{PV}(h) = \sum_{i=1}^{N}(y_i - \hat{m}(x_i))^2\pi(x_i)p(w_{ii,h})$$

其中，$p(\cdot)$ 表示惩罚函数，而 $w_{ii,h}$ 表示 $\hat{m}(x_i)=\sum_j w_{ji,h}y_j$ 中给定第 i 个观测值时的权数。

一个广为流行的例子是广义交叉验证测量(**generalized cross-validation measure**)，它使用惩罚函数 $p(w_{ii,h})=(1-w_{ii,h})^2$。其他一些惩罚函数已由哈德尔(Härdle, 1990，第167页)以及哈德尔和林顿(Härdle and Linton, 1994，第2 323页)给出。

交叉验证例子

就9.4.2节中局部进行平均的例子而言，对于 $k=3, 5, 7, 9, 25$，CV(k)=54 811, 56 666, 63 456, 65 605, 69 939。在这种情况下，所有观测值都用于计算 CV(k)，满足 $\pi(x_i)=1$，尽管可能出现端点问题。在 $k=5$ 之后，没有实际提高，即使从图9.5中看，这个值产生太粗略的估计值，但在实际应用时，人们愿意选取比

较大的 k 值，以便得到较为光滑的曲线。

更一般地讲，交叉验证绝不是完美无缺的，为了获得人们所期望的光滑程度，一种普遍方法是采用“目测”选取 h 来拟合非参数曲线。

修饰

式(9.21)核估计量的分母是 $\hat{f}(x_0)$，即回归元的密度在 x_0 点的核估计值。在一些计算点上，$\hat{f}(x_i)$ 可以非常小，从而导致非常大的估计值 $\hat{m}(x_i)$。修饰[1]（**trimming**）可消除或大大降低满足 $\hat{f}(x_i)<b$ 所有点的权数，比如说当 $N\to\infty$ 时，b 以适当速率 $b\to 0$。这类问题可能在分布尾部出现。对于非参数估计来说，人们仅仅关注于 x_i 的更居中心值的 $m(x_i)$ 估计，以及尾部中可能在交叉验证降低权数的那些值。不过，9.7 节的半参数方法必须在 x_i 的所有值上计算，在此情况下，进行修饰就不足为奇了。从原则上讲，尽管在有限样本上修饰将会有些差异，但在渐近形式上修饰函数应该没有差异。

9.5.4 置信区间

通常，核回归估计应以逐点置信区间加以阐述。一种简单方法是，阐述在 x_0 点处计算的 $f(x_0)$ 的逐点置信区间，例如 x_0 等于 x 的第 1 个十分位数直到第 9 个十分位数。

当忽略 $\hat{m}(x_0)$ 中的偏倚，由式(9.24)知，得到下述 95%置信区间（**confidence interval**）：

$$m(x_0)\in\hat{m}(x_0)\pm 1.96\sqrt{\frac{1}{Nh}\frac{\hat{\sigma}_\varepsilon^2}{\hat{f}(x_0)}\int K(z)^2dz}$$

其中，$\hat{\sigma}_\varepsilon^2=\sum_i w_{i0,h}\hat{\varepsilon}_i^2$，而 $w_{i0,h}$ 已由式(9.22)定义，$\hat{f}(x_0)$ 表示在 x_0 点的核密度估计。该估计假定同方差误差，尽管对异方差性可能是稳健的，因为接近 x_0 的观测值被赋予最大权数。否则，由式(9.20)之后的讨论知，异方差稳健的 95%置信区间是 $\hat{m}(x_0)\pm 1.96\hat{s}_0$，其中，$\hat{s}_0^2=\sum_i w_{i0,h}^2\hat{\varepsilon}_i^2$。

如同在核密度情况下一样，$\hat{m}(x_0)$ 中的偏倚不应被忽略。正如已注意到的，对偏倚进行估计很困难。然而，一种标准方法是，就较小带宽 h 而言，满足 $h=o(N^{-0.2})$ 而不是最优的 $h^*=O(N^{-0.2})$，进行光滑不足。

哈德尔(Härdle，1990)曾经给出详细的置信区间表述，包括一致置信带而不是逐点区间，而自助法则将在 11.6.5 节加以详述。

9.5.5 导数估计

当进行回归时，我们经常对 y 的条件均值如何随 x 变化而变动感兴趣，即边际效应（**marginal effect**）而不是条件均值本身。

很容易用核估计求导数。一个一般性结果是，核回归估计的第 s 阶导数 $\hat{m}^{(s)}(x_0)$ 关于 $m^{(s)}(x_0)$ 是一致的，$m^{(s)}(x_0)$ 表示条件均值 $m(x_0)$ 的第 s 阶导数，人们

[1] 又称为修剪。——译者注

能采用微分法,或者采用有限差分法加以计算。

举一个例子,考虑前一节的数据生成例子中一阶导数的估计。设 $z_1,\cdots,z_N$ 表示有序点,核回归函数在这些点上进行计算,而 $\hat{m}(z_1),\cdots,\hat{m}(z_N)$表示在这些点上的估计值。有限差分估计是 $\hat{m}'(z_i)=[\hat{m}(z_i)-\hat{m}(z_{i-1})]/[z_i-z_{i-1}]$。由式(9.17)给出的数据生成过程(dgp)是二次形式 $\hat{m}'(z_i)=6.5-0.30z_i+0.003z_i^2$,图 9.7 画出有限差分估计以及实际导数。正如人们所料,导数估计有点繁琐,但它却能抓住本质。导数估计应建立在对条件均值过度光滑估计的基础上。对于更详细内容,参见帕甘和乌拉(Pagan and Ullah, 1999, 第 4 章)。哈德尔(Härdle, 1990, 第 160 页)阐述了交叉验证对导数估计的修改。

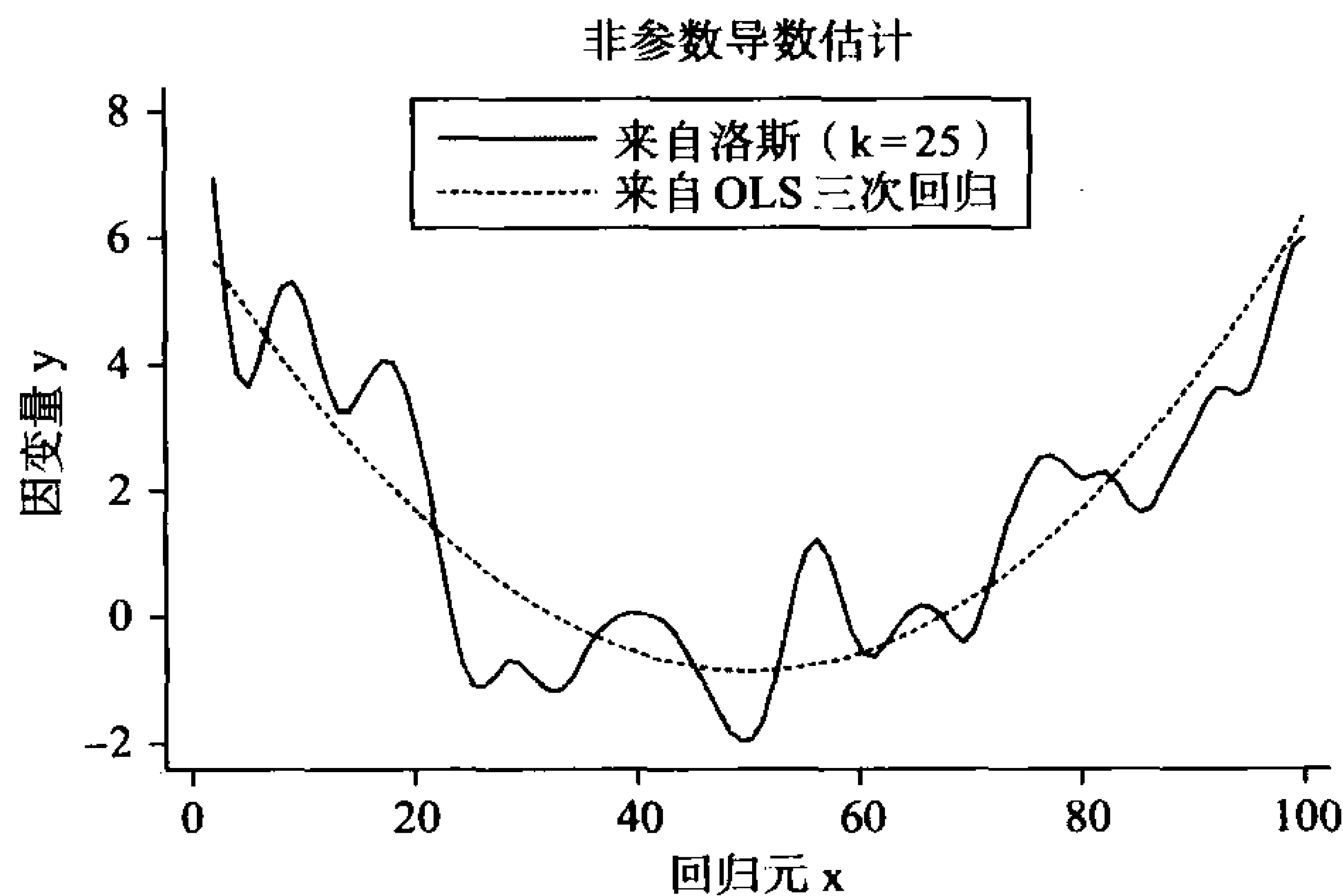


图 9.7 利用前面估计的洛斯回归曲线与三次回归曲线的非参数导数估计。数据生成过程与图 9.5 的一样。

除局部导数 $m'(x_0)$之外,我们还对平均导数 $\mathrm{E}[m'(x)]$感兴趣。由 9.7.4 节给出的平均导数估计量提供了$\sqrt{N}$一致的且渐近正态的 $\mathrm{E}[m'(x)]$估计。

9.5.6 *条件矩估计*

条件均值 $\mathrm{E}[y|x]=m(x)$的核回归方法能够被推广到其他条件矩的非参数估计上。

对于原条件矩(**conditional moments**)譬如 $\mathrm{E}[y^k|x]$来说,我们用加权平均:

$$\hat{\mathrm{E}}[y^k|x_0]=\sum_{i=1}^{N}w_{i0,h}y_i^k \tag{9.28}$$

其中,权数 $w_{i0,h}$与估计 $m(x_0)$时所用的权数一样。

于是,中心条件矩能通过把它们重新表达成原始矩的加权和而得以计算。例如,由于 $\mathrm{V}[y|x]=\mathrm{E}[y^2|x]-(\mathrm{E}[y|x])^2$,所以通过 $\hat{\mathrm{E}}[y^2|x_0]-\hat{m}(x_0)^2$ 来估计其条件方差。人们发现,与对条件均值估计相比,对较高阶的条件矩进行估计更为繁琐。

9.5.7 *多元变量核回归*

前面讨论了单个回归元的核回归。对于纯量 y 对 k 维向量 $\mathbf{x}$ 的回归即 $y_i=m(\mathbf{x}_i)+\varepsilon_i=m(x_{1i},\cdots,x_{1k})+\varepsilon_i$ 来说,$m(\mathbf{x}_0)$的核估计量变成:

$$\hat{m}(\mathbf{x}_0) \equiv \frac{\frac{1}{Nh^k}\sum_{i=1}^{N} K\left(\frac{\mathbf{x}_i - \mathbf{x}_0}{h}\right) y_i}{\frac{1}{Nh^k}\sum_{i=1}^{N} K\left(\frac{\mathbf{x}_i - \mathbf{x}_0}{h}\right)}$$

其中，$K(\cdot)$现在表示多元变量核(**multivariate kernel**)。尽管使用多元变量核，譬如多元变量正态密度，但 $K(\cdot)$经常是 k 个一维核之积。

当使用乘积核时，回归元应通过除以标准差变换成为一种共同标度。然后，交叉验证测量式(9.25)能用于决定共同的最优带宽 h^*，虽然要决定哪一个 $\mathbf{x}_i$ 应该被降低权数，因为当 $\mathbf{x}$ 是多变量时，接近于端点的闭性结果就更为复杂。否则，回归元无须重新标度，然而，对于每一个回归元都应该使用各种不同的带宽。

由于估计又一次是 y_i 的一种局部平均，所以渐近结果与表达式均类似于前面曾考察的那些结果。如同前面一样，偏倚 $b(\mathbf{x}_0)$再次是 $O(h^2)$，但 $\hat{m}(\mathbf{x}_0)$的方差却以速率 $O(Nh^k)$下降，它比一维情况收敛得更慢，因为实际上样本中很小部分被用于求 $\hat{m}(\mathbf{x}_0)$。于是：

$$\sqrt{Nh^k}(\hat{m}(\mathbf{x}_0) - m(\mathbf{x}_0) - b(\mathbf{x}_0)) \xrightarrow{d} \mathcal{N}\left[0, \frac{\sigma_\varepsilon^2}{f(\mathbf{x}_0)}\int K(\mathbf{z})^2 d\mathbf{z}\right]$$

最优带宽选取是 $h^* = O(N^{-1/(k+4)})$，这比一维情况要大一些。相应的 $\hat{m}(\mathbf{x}_0)$最优速率是 $N^{-2/(k+4)}$。

这一结果与前面的一些纯量结果均假定，$m(x)$是二次可微的，即获得式(9.23)偏倚项的必要条件，可是当 $m(x)$是 p 次可微时，利用 p 阶有序核来使偏倚的阶数减少(参见 9.3.3 节)，从而导致较小的 h^* 并且达到 9.4.5 节给出的斯通界(**Stone's bound**)的较快收敛速率；参见哈德尔(Härdle, 1990，第 93 页)的更详细内容。下一节给出的其他非参数估计量也能达到斯通界。

随着回归元个数增加，收敛速率会减少，而当回归元个数趋向于无穷时，收敛速率趋于 N^0。这种维数祸根(**curse of dimensionality**)大大限制了具有几个回归元的回归模型中对非参数方法的使用。半参数模型(参见 9.7 节)设置了额外结构，以使非参数成分具有很小维数。

9.5.8 参数模型检验

对条件均值参数模型进行正确设定的一种明显检验是，把拟合均值与从非参数模型中获得的值进行比较。

设 $\hat{m}_\theta(\mathbf{x})$表示 $\mathrm{E}[y|\mathbf{x}]$的参数估计量，而 $\hat{m}_h(\mathbf{x})$表示非参数估计量，譬如核估计。一种方法是在 $\mathbf{x}$ 值范围内，把 $\hat{m}_\theta(\mathbf{x})$与 $\hat{m}_h(x)$进行比较。这因为需要 $\hat{m}_\theta(\mathbf{x})$中正确的渐近偏倚而变得复杂[参见哈德尔和玛门(Härdle and Mammen, 1993)]。第二种方法是，考察形式为 $N^{-1}\sum_i w_i(y_i - \hat{m}_\theta(\mathbf{x}_i))$的条件矩检验，其中，各种不同权数部分地建立在核回归基础上，这些不同权数用于检验各种不同方向的 $\mathrm{E}[y|\mathbf{x}] = \hat{m}_\theta(\mathbf{x})$的成立与否。例如，霍罗维茨和哈德尔(Horowitz and Härdle, 1994)使用了 $w_i = \hat{m}_h(\mathbf{x}_i) - \hat{m}_\theta(\mathbf{x}_i)$。帕甘和乌拉(Pagan and Ullah, 1999，第 141～150 页)及亚特丘(Yatchew, 2003，第 119～124 页)对所使用的方法给出了一个综述。

9.6 可供选择的非参数回归估计量

9.4 节曾引入局部回归方法，即该方法通过局部加权平均 $\hat{m}(x_0)=\sum_i w_{i0,h} y_i$ 估计回归函数 $m(x_0)$，其中，权数 $w_{i0,h}=w(x_i,x_0,h)$ 不同于 x_0 点处的计算值，但与 x_i 点处的值相同。9.5 节已经阐述了权数都是核权数时的详细结果。

这里，我们考察对应于其他权数的一些普遍使用的估计量。尽管关于偏倚与方差的准确表达式不同于式(9.23)与式(9.24)中的那些结果，但是对于使用类似的最优收敛率与带宽选取的交叉验证，9.5 节中的许多结果都可以完成。9.6.2 节给出的估计量尤其流行。

9.6.1 *最近邻估计量*

k-最近邻估计量(**k-nearest neighbor estimator**)是最接近 x_0 的 k 个 x_i 观测值的那些 y 值的等权平均。将 $N_k(x_0)$ 定义成最接近 x_0 的 k 个 x_i 观测值的集合。于是：

$$\hat{m}_{k\text{-}NN}(x_0)=\frac{1}{k}\sum_{i=1}^{N}\mathbf{1}(x_i\in N_k(x_0))y_i \tag{9.29}$$

该估计量是带有一致权数的核估计量(参见表 9.1)，只是其带宽变化。此处，在 x_0 点的带宽 h_0 等于 x_0 与 k 个最近邻中的最远者之间的距离，而且更正式地，有 $h_0\simeq k/(2Nf(x_0))$。数量 k/N 称为跨距[1](**span**)。比较光滑的曲线可利用式(9.29)中的核权数来获得。

这个估计量因为提供了可变带宽选择的简单规则而引人注目。从计算上看，一种较快的方式是，使用对称形式(**symmetrized**)，即使用 $k/2$ 个左边的最近邻以及相同个数右边的最近邻，这是 9.4.2 节用过的局部平均方法。从而，人们能运用依 x_i 递增顺序而排列的观测值校正公式，从而当 x_0 增大时，一个观测值离开数据，而另一个观测值进入数据。

9.6.2 *局部线性回归与洛斯回归*

核回归估计量是一种局部常值估计量(**local constant estimator**)，因为它假定 $m(x)$ 在 x_0 的局部邻域之内为常值。可是，人们能设 $m(x)$ 在 x_0 邻域之内是线性的，因而在 x_0 某个邻域中，$m(x)=a_0+b_0(x-x_0)$。

为了实施这一想法，注意到，核回归估计量 $\hat{m}(x_0)$ 可通过对 $\sum_i K((x-x_0)/h)\times(y_i-m_0)^2$ 求关于 m_0 的极小值而获得。局部线性回归估计量(**local linear regression estimator**)求：

$$\sum_{i=1}^{N}K\left(\frac{x_i-x_0}{h}\right)(y_i-a_0-b_0(x_i-x_0))^2 \tag{9.30}$$

[1] 又称为支点距。——译者注

关于 a_0 与 b_0 的极小值，其中，$K(\cdot)$ 表示核加权函数。于是，在 x_0 邻域内，$\hat{m}(x)=\hat{a}_0+\hat{b}_0(x-x_0)$。然后，在 x_0 点处，估计值准确地是 $\hat{m}(x)=\hat{a}_0$，而 $\hat{b}_0$ 提供了一阶导数 $\hat{m}'(x_0)$ 的估计值。更一般地讲，p 次局部多项式估计量(**local polynomial estimator of degree**)求：

$$\sum_{i=1}^{N} K\left(\frac{x_i-x_0}{h}\right)\left(y_i-a_{0,0}-a_{0,1}(x_i-x_0)-\cdots-a_{0,p}\frac{(x_i-x_0)^p}{p!}\right)^2 \quad (9.31)$$

的极小值，得到 $\hat{m}^{(s)}(x_0)=\hat{a}_{0,s}$。

范剑青和吉贝尔斯(Fan and Gijbels, 1996)对该方法给出了许多性质，并阐明其引人注目的原因。在每一个计算点 x_0 上，估计仅仅需要加权最小二乘法回归。此估计量能表述成 y_i 的一个加权平均，因为它们都是 LS 估计量。局部线性估计量具有偏倚项 $b(x_0)=h^2\left(\frac{1}{2}m''(x_0)\right)\int z^2K(z)dz$，与式(9.23)给出的核回归偏倚不一样，$b(x_0)$ 不依赖于 $m'(x_0)$。这特别有助于克服 9.4.2 节已经阐述的边界问题。为了估计第 s 阶导数，对 p 的一个好选取是 $p=s+1$，因此，例如，人们使用局部二次估计量来估计一阶导数。

标准的局部回归估计量是局部加权散点光滑法(**locally weighted scatterplot smoothing**，简记为 LOWESS)或克利夫兰(Cleveland，1979)的洛斯估计量(**Lowess estimator**)。这是局部多项式估计的一种变形，局部多项式估计式(9.31)使用了由从 x_0 到 x_0 的第 k 个最近邻点距离决定的可变带宽 $h_{0,k}$，使用了 9 次核 $K(z)=(70/81)(1-|z|^3)^3\mathbf{1}(|z|<1)$；并对具有大残差 $y_i-\hat{m}(x_i)$ 的观测值降低加权，这需要经过数据 N 次。有关综述，参见范剑青和吉贝尔斯(Fan and Gijbels，1996，第 24 页)。与核回归相比，洛斯(Lowess)估计量则引人注目，因为它使用可变带宽，对离群值来说是稳健的，并利用局部多项式估计量求边界问题的极小值。可是，它是一种密集计算。

另外一种流行的变形是，弗里德曼(Friedman，1984)的超光滑子(**supersmoother**)[参见哈德尔(Härdle,1990,第 181 页)]。为了更好地在边界处利用局部线性拟合而不是局部常值拟合，一个起点是对称的 k - NN。可是，超光滑子是一种可变跨距光滑子(**smoother**)而不是用固定跨距或固定的 k，其中，可变跨距是由局部交叉验证来确定的，交叉验证需要 9 次转移数据。与洛斯估计量相比，超光滑子对离群值来说不是稳健的，但它却允许跨距变化且是快速计算的。

9.6.3 *光滑样条估计量*

三次光滑样条估计量(**cubic smoothing spline estimator**)$\hat{m}_\lambda(x)$ 对惩罚残差平方和：

$$\mathrm{PRSS}(\lambda)=\sum_{i=1}^{N}(y_i-m(x_i))^2+\lambda\int(m''(x))^2dx \quad (9.32)$$

求极小值，其中，λ 表示光滑系数。如同本章其他地方一样，使用平方误差损失。第一项只会产生相当粗糙的拟合，进而 $\hat{m}(x_i)=y_i$。第二项引入惩罚粗糙度

(**roughness**)。9.5.3 节的交叉验证方法可用于确定 λ,对于 λ 较大值来说,会导致较光滑的曲线。

哈德尔(Härdle, 1990, 第 56～65 页)已经证明,$\hat{m}_\lambda(x)$ 关于逐次 x 值是三项多项式的,而且此估计量能表达成 y 值的一个局部加权平均,并且是渐近等价于具有特殊可变核的核估计量。在微观经济计量学中,与本章其他方法相比,光滑样条并不经常使用。此方法适用于其他的粗糙度惩罚与其他的损失函数。

9.6.4 *序列估计量*

序列估计量是通过 K 个函数 $z_1(x),\cdots,z_K(x)$ 的加权和逼近回归函数:

$$\hat{m}_K(x)=\sum_{j=1}^{K}\hat{\beta}_j z_j(x) \tag{9.33}$$

其中,系数 $\hat{\beta}_1(x),\cdots,\hat{\beta}_K$ 可直接通过 y 对 $z_1(x),\cdots,z_K(x)$ 的 OLS 回归获得。函数 $z_1(x),\cdots,z_K(x)$ 构成了一个截取序列。一些例子包括,第 $(K-1)$ 阶多项式逼近或者满足 $z_j(x)=x^{j-1}$ 的幂序列 $j=1,\cdots,K$;正交变形以及标准正交多项式变形(参见 12.3.1 节);截取傅里叶序列,其中回归元被重新标度,从而 $x\in[0,2\pi]$;加伦特(Gallant, 1981)的傅里叶灵活函数形式,它是一种截取傅里叶序列加上 x 与 x^2 项;通过在给定结点(knots)个数之间的多项式函数逼近回归函数 $m(x)$ 的回归样条,这些函数在结点处连接在一起。

此方法不同于 9.4 节中的方法,因为它是估计 $m(x)$ 的一种全局近似方法。不过,如果当 $N\to\infty$ 时,以适当速率 $K\to\infty$,那么 $\hat{m}_K(x)\xrightarrow{p}m(x_0)$。由纽韦(Newey, 1997)知,如果 $\mathbf{x}$ 是 k 维的且 $m(\mathbf{x})$ 是 p 次可微的,那么积分均方误差(参见 9.5.3 节)$\text{MISE}(h)=O(K^{-2p/k}+K/N)$,其中,第一项反映偏倚,而第二项则反映方差。令这些式子相等,得出最优 $K^*=N^{k/(2p+k)}$,所以 K 增大,却以比样本量较低的速率增长。$\hat{m}(x)$ 的收敛速率等于 9.4.5 节给出的斯通(Stone, 1980)最快可能速率。

从直观上讲,序列估计量并不是稳健的,因为离散值可能是全局性的而不仅仅为局部影响 $\hat{m}(x)$,但这一猜想在教科书给出的典型例子中无须检验。

安德鲁斯(Andrews, 1991)与纽韦(Newey, 1997)给出一个包括多元变量情况的非常一般性的研究,研究内容包括泛函估计而不是条件均值,以及对半参数模型的推广,其中序列方法是最经常使用的。

9.7 半参数回归

上述分析,在没有任何结构情况下强调了回归模型。在微观经济计量学中,通常把某种结构施加到回归模型上。

首先,在需求函数中,经济理论会施加某种结构,譬如对称性与同质性。这类信息会被并入非参数回归中;例如,参见马茨金(Matzkin, 1994)。

其次,也更为广泛出现的,经济计量模型包括众多潜在回归元,以致维数祸根完全使得非参数分析不切合实际。然而,一种普遍方法是估计半参数模型

(**semiparametric model**),粗略地讲,半参数回归是把参数成分与非参数成分结合起来,参见鲍威尔(Powell, 1994)对半参数术语的详细讨论。

存在许多不同的半参数模型,而且相当多的方法均可用于一致地估计这些模型。在本节,我们仅仅阐述几个重要例子。一些应用也会在本书的其他地方给出,包括第 14 章和第 16 章给出的二值结果模型与删失回归模型。

9.7.1 *例子*

表 9.2 已列出半参数回归的几个重要例子。前两个例子通过增加未设定成分 $\lambda(\mathbf{z})$,或者通过允许未设定变换 $g(\mathbf{x}'\boldsymbol{\beta})$ 来推广线性模型 $\mathbf{x}'\boldsymbol{\beta}$,这将在下面详细介绍,而第三个例子是对前两个例子的组合。接下来的三个例子在应用统计学中的应用比经济计量学的应用更为广泛,它们是通过假定回归元的可加性或可分性来减少维数,否则就是非参数模型。我们将详述广义可加模型。与这些模型有关的是神经网络模型(**neural network models**);参见库安和怀特(Kuan and White, 1994)。最后一个例子是条件方差的一种灵活模型,下面也将对此模型加以详述。为确保半参数模型是可识别的,需要小心谨慎地处理。例如,参见单指标模型的讨论。除估计 $\boldsymbol{\beta}$ 之外,关注内容还在于边际效应上,比如 $\partial \mathrm{E}[y|\mathbf{x},\mathbf{z}]/\partial \mathbf{x}$。

表 9.2 半参数模型:一些重要例子

名称	模型	参数	非参数
偏线性	$\mathrm{E}[y\|\mathbf{x},\mathbf{z}]=\mathbf{x}'\boldsymbol{\beta}+\lambda(\mathbf{z})$	$\boldsymbol{\beta}$	$\lambda(\cdot)$
单指标	$\mathrm{E}[y\|\mathbf{x}]=g(\mathbf{x}'\boldsymbol{\beta})$	$\boldsymbol{\beta}$	$g(\cdot)$
广义偏线性	$\mathrm{E}[y\|\mathbf{x},\mathbf{z}]=g(\mathbf{x}'\boldsymbol{\beta}+\lambda(\mathbf{z}))$	$\boldsymbol{\beta}$	$g(\cdot)$, $\lambda(\cdot)$
广义可加的	$\mathrm{E}[y\|\mathbf{x}]=c+\sum_{j=1}^{k}g_j(x_j)$	—	$g_j(\cdot)$
偏可加的	$\mathrm{E}[y\|\mathbf{x},\mathbf{z}]=\mathbf{x}'\boldsymbol{\beta}+c+\sum_{j=1}^{k}g_j(z_j)$	$\boldsymbol{\beta}$	$g_j(\cdot)$
投影寻踪	$\mathrm{E}[y\|\mathbf{x}]=\sum_{j=1}^{M}g_j(\mathbf{x}_j'\boldsymbol{\beta}_j)$	$\boldsymbol{\beta}_j$	$g_j(\cdot)$
异方差线性的	$\mathrm{E}[y\|\mathbf{x}]=\mathbf{x}'\boldsymbol{\beta}$; $\mathrm{V}[y\|\mathbf{x}]=\sigma^2(\mathbf{x})$	$\boldsymbol{\beta}$	$\sigma^2(\cdot)$

9.7.2 *半参数估计量的效率*

在阐述几个重要例子的半参数模型结果之前,我们考察通过半参数方法而不是参数方法进行估计时有效性的损失。

我们在这里的概述沿着鲁宾逊(Robinson, 1998b)的线索,他曾考察具有参数成分(记为 $\boldsymbol{\beta}$)与非参数成分(记为 G)的半参数模型,非参数成分 G 依赖于无限多个冗余参数。G 的一些例子包括服从对称分布的 iid 误差分布形状,以及在 9.7.4 节将由式(9.7.3)给出的单指标函数 $g(\cdot)$。估计量 $\hat{\boldsymbol{\beta}}=\boldsymbol{\beta}(\hat{G})$,其中,$\hat{G}$ 表示 G 的非参数估计量。

原则上讲,一个估计量 $\hat{\boldsymbol{\beta}}$ 是适应的(**adaptive**),是指在通过非参数方法估计 G 时无有效性损失,所以:

$$\sqrt{N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\xrightarrow{d}\mathcal{N}[\mathbf{0},\mathrm{V}_G]$$

其中，V_G 表示所考虑的特定类中任何形状函数 G 的协方差矩阵。在似然框架下，V_G 是克莱默—劳(Cramer - Rao)下界。在二阶矩背景下，V_G 是由高斯—马尔可夫定理或推广形式譬如 GMM 得出的。适应估计量的一个重要例子是，含有设定条件均值函数的且含有异方差性的未知函数形式的估计。

理论上，如果估计量 $\hat{\beta}$ 不是适应的，那么接下来的最佳最优性质会使该估计量达到半参数有效界(**semiparametric efficiency bounds**)V_G^*，所以：

$$\sqrt{N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}[\mathbf{0}, V_G^*]$$

其中，V_G^* 表示克莱默—劳下界的推广或者它的二阶矩类似形式，该二阶矩类似形式提供了给定设定半参数模型时可能的最小方差矩阵。对于适应估计量来说，$V_G^*=V_G$，但是通常 V_G^* 大于 V_G。半参数有效性界将在 9.7.8 节引入。它们仅仅在某些半参数设置下能获得，并且甚至当它们是已知的时候，不存在任何一个达到此界的估计量。达到此界的一个例子是，克莱因和斯帕迪(Klein and Spady, 1993)的二值选择模型估计量(参见 14.7.4 节)。

倘若半参数有效性界没有达到或不是已知的，则接下来的最佳性质是，当 V_G^{**} 大于 V_G^* 时，$\sqrt{N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}[\mathbf{0}, V_G^{**}]$，这使进行通常统计推断成为可能。更一般地讲，$\sqrt{N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})=O_p(1)$，但不必是正态分布。最后，一致的但小于$\sqrt{N}$一致的估计量具有性质 $N^r(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})=O_p(1)$，其中，$r<0.5$。通常，不能得到渐近正态分布。当对参数与非参数部分同等处理时，就出现这种情况，因而在 $\boldsymbol{\beta}$ 与 G 上共同达到极大化。存在许多例子，尤其是在离散选择模型与截取选择模型之中。

尽管半参数估计量具有潜在无效性，但它们仍是引人注目的，因为在完全参数估计量是非一致的背景下，半参数估计量仍保持一致性。鲍威尔(Powell, 1994)给出一张表格，并阐述一系列半参数模型的一致性存在，以及$\sqrt{N}$一致渐近正态估计量的归纳总结。

9.7.3 *偏线性模型*

偏线性模型[1](**partially linear model**)是将条件均值设定成为通常线性回归函数加一个未设定的非线性成分，因此有：

$$\mathrm{E}[y|\mathbf{x},\mathbf{z}]=\mathbf{x}'\boldsymbol{\beta}+\lambda(\mathbf{z}) \tag{9.34}$$

其中，纯量函数 $\lambda(\cdot)$ 表示未设定的。

一个例子是需求函数关于弹性的估计，其中，$\mathbf{z}$ 反映出一天时间或天气指示变量，诸如温度。第二个例子是 16.5 节给出的样本选择模型。由于省略变量偏倚的缘故，所以一旦忽略 $\lambda(\mathbf{z})$，就会导致非一致的 $\boldsymbol{\beta}$，除非 $\mathrm{Cov}[\mathbf{x},\lambda(\mathbf{z})]=\mathbf{0}$。在一些应用中，关注内容在于 $\boldsymbol{\beta}$、$\lambda(\mathbf{z})$或者两者都有。对 $\mathrm{E}[y|\mathbf{x},\mathbf{z}]$ 进行完全非参数估计是可行的，却会使 $\boldsymbol{\beta}$ 出现小于$\sqrt{N}$一致估计的情况。

---

[1] 又称为部分线性模型。——译者注

鲁宾逊差分估计量

不过，鲁宾逊(Robinson，1988a)曾经提出下述方法。回归模型蕴含：

$$y=\mathbf{x}'\boldsymbol{\beta}+\lambda(\mathbf{z})+u$$

其中，误差 $u=y-\mathrm{E}[y|\mathbf{x},\mathbf{z}]$。反之，这蕴含：

$$\mathrm{E}[y|\mathbf{z}]=\mathrm{E}[\mathbf{x}|\mathbf{z}]'\boldsymbol{\beta}+\lambda(\mathbf{z})$$

这是因为 $\mathrm{E}[u|\mathbf{x},\mathbf{z}]=0$ 蕴含 $\mathrm{E}[u|\mathbf{z}]=0$。一旦对这两个式子相减，得到：

$$y-\mathrm{E}[y|\mathbf{z}]=(\mathbf{x}-\mathrm{E}[\mathbf{x}|\mathbf{z}])'\boldsymbol{\beta}+u \tag{9.35}$$

式(9.35)中的条件矩是未知的，但它们却能用非参数估计代替。

因而，鲁宾逊提出：

$$y_i-\hat{m}_{yi}=(\mathbf{x}-\hat{\mathbf{m}}_{\mathbf{x}i})'\boldsymbol{\beta}+v \tag{9.36}$$

的 OLS 估计，其中，$\hat{m}_{yi}$ 与 $\hat{\mathbf{m}}_{\mathbf{x}i}$ 分别表示来自 y_i 与 $\mathbf{x}_i$ 对 $\mathbf{z}_i$ 的非参数回归的预测值。给定关于 i 的独立性，假定 u_i 是 iid $[0,\sigma^2]$ 的，式(9.36)中 $\boldsymbol{\beta}$ 的 OLS 估计量是 $\sqrt{N}$ 一致的且渐近正态的，满足：

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{\mathrm{PL}}-\boldsymbol{\beta})\xrightarrow{d}\mathcal{N}\left[\mathbf{0},\sigma^2\left(\mathrm{plim}\,\frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i-\mathrm{E}[\mathbf{x}_i|\mathbf{z}_i])(\mathbf{x}_i-\mathrm{E}[\mathbf{x}_i|\mathbf{z}_i])'\right)^{-1}\right]$$

不对 $\lambda(\mathbf{z})$ 进行设定，会导致有效性损失，尽管当 $\mathrm{E}[\mathbf{x}|\mathbf{z}]$ 关于 $\mathbf{z}$ 是线性的时，就没有损失。为了估计 $\mathrm{V}[\hat{\boldsymbol{\beta}}_{\mathrm{PL}}]$，直接用 $(\mathbf{x}_i-\hat{\mathbf{m}}_{\mathbf{x}i})$ 代替 $(\mathbf{x}_i-\mathrm{E}[\mathbf{x}_i|\mathbf{z}_i])$。其渐近结论能推广到异方差性误差上，在此情况下，人们刚好使用源自 OLS 回归(9.36)的通常艾克—怀特标准误差。由于 $\lambda(\mathbf{z})=\mathrm{E}[y|\mathbf{z}]-\mathrm{E}[\mathbf{x}|\mathbf{z}]'\boldsymbol{\beta}$，所以它可由 $\hat{\lambda}(z)=\hat{m}_{yi}-\hat{\mathbf{m}}_{\mathbf{x}i}'\hat{\boldsymbol{\beta}}$ 一致地加以估计。

人们能使用各种非参数估计量 $\hat{m}_{yi}$ 与 $\hat{\mathbf{m}}_{\mathbf{x}i}$。鲁宾逊(Robinson，1988a)使用要求收敛速率不低于 $N^{-1/4}$ 的核估计，因此，当 $\mathbf{z}$ 的维数很大时，就需要过度光滑的或者较高阶的核；参见帕甘和乌拉(Pagan and Ullah，1999，第 205 页)。还注意到，核估计量可能需要加以修饰(参见 9.5.3 节)。

其他估计量

在偏线性模型中，几种其他方法会得到 $\boldsymbol{\beta}$ 的 $\sqrt{N}$ 一致估计值。斯佩克曼(Speckman，1988)还曾经用过核。恩格尔等人(Engle et al.，1986)使用三次光滑样本估计量的推广。安德鲁斯(Andrews，1991)阐述了 y 对 $\mathbf{x}$ 的回归以及 9.6.4 节给出的关于 $\lambda(\mathbf{z})$ 的序列近似。亚特丘(Yatchew，1997)阐述了简单差分估计量。

9.7.4 *单指标模型*

单指标模型(**single-index model**)是将条件均值设定成回归元线性组合的一种未知纯量函数，满足：

$$\mathrm{E}[y|\mathbf{x}]=(\mathbf{x}'\boldsymbol{\beta}) \tag{9.37}$$

其中，纯量函数 $g(\cdot)$ 是未设定的。单指标模型的优点已在 5.2.4 节阐述过。这里的

函数$g(\cdot)$可从数据中获得，不过，在前面一些例子中，则设定 $\mathrm{E}[y|\mathbf{x}]=\exp(\mathbf{x}'\boldsymbol{\beta})$。

识别

市村(Ichimura，1993)已经阐述单指标模型的识别条件(**identification conditions**)。对于未知函数 $g(\cdot)$来说，单指标模型 $\boldsymbol{\beta}$是可识别的，至多仅差一个位置与标度。为了理解这一点，注意到，就纯量 v 而言，函数 $g^*(a+bv)$总能表述成$g(v)$，因而函数 $g^*(a+b\mathbf{x}'\boldsymbol{\beta})$等价于 $g(\mathbf{x}'\boldsymbol{\beta})$。此外，$g(\cdot)$必须是可微的。在最简单情况下，所有估计量都是连续的。相反，如果某些回归元是离散的，那么至少有一个回归元必须是连续的，而且当 $g(\cdot)$是单调函数时，就能达到 β 的界。

平均导数估计量

对于连续回归元来说，斯托克(Stoker，1986)曾经发现，如果条件均值是单指标的，那么条件均值的平均导数向量就能确定 β，至多仅差一个标度而已，因为就 $m(\mathbf{x}_i)=g(\mathbf{x}_i'\boldsymbol{\beta})$而言，有：

$$\boldsymbol{\delta}\equiv \mathrm{E}\left[\frac{\partial m(\mathbf{x})}{\partial \mathbf{x}}\right]=\mathrm{E}[g'(\mathbf{x}'\boldsymbol{\beta})]\boldsymbol{\beta} \tag{9.38}$$

而 $\mathrm{E}[g'(\mathbf{x}_i'\boldsymbol{\beta})]$是一个纯量。进一步地，由 5.6.3 节给出的广义信息矩阵等式知，对于任何函数 $h(\mathbf{x})$，$\mathrm{E}[\partial h(\mathbf{x})/\partial\mathbf{x}]=-\mathrm{E}[h(\mathbf{x})s(\mathbf{x})]$，其中，$s(\mathbf{x})=\partial\ln f(\mathbf{x})/\partial\mathbf{x}=f'(\mathbf{x})/f(\mathbf{x})$，而 $f(\mathbf{x})$表示 $\mathbf{x}$ 的密度。因而：

$$\boldsymbol{\delta}=-\mathrm{E}[m(\mathbf{x})s(\mathbf{x})]=-\mathrm{E}[\mathrm{E}[y\mid\mathbf{x}]s(\mathbf{x})] \tag{9.39}$$

由此可得，通过平均导数估计量[**average derivative (AD) estimator**]：

$$\hat{\boldsymbol{\delta}}_{\mathrm{AD}}=-\frac{1}{N}\sum_{i=1}^{N}y_i\hat{s}(\mathbf{x}_i) \tag{9.40}$$

能估计 $\boldsymbol{\delta}$，从而估计出 $\boldsymbol{\beta}$，只是至多差一个标度，其中，$\hat{s}(\mathbf{x}_i)=\hat{f}'(\mathbf{x}_i)/\hat{f}(\mathbf{x}_i)$能通过 $\mathbf{x}_i$ 密度的核估计及其一阶导数得到估计。该估计量 $\hat{\boldsymbol{\delta}}$ 是$\sqrt{N}$一致的，而且其渐近正态分布已经由哈德尔和斯托克(Härdle and Stoker，1989)推导出。通过 y_i 对 $\mathbf{x}_i'\hat{\boldsymbol{\delta}}$ 的非参数回归，估计出函数 $g(\cdot)$。注意到，不管单指标模型是否有联系，$\hat{\boldsymbol{\delta}}_{\mathrm{AD}}$都给出 $\mathrm{E}[m'(\mathbf{x})]$的一个估计值。

$\hat{\boldsymbol{\delta}}_{\mathrm{AD}}$的一个弱点是，如果 $\hat{f}(\mathbf{x}_i)$很小，则 $\hat{s}(\mathbf{x}_i)$非常大。一种可能性是，当 $\hat{f}(\mathbf{x}_i)$很小时，就要进行修饰。不过，鲍威尔、斯托克和斯托克(Powell，Stock，and Stoker，1989)发现，结果式(9.38)可推广到含有 $\boldsymbol{\delta}\equiv\mathrm{E}[w(\mathbf{x})m'(\mathbf{x})]$的加权导数上。特别地，选取 $w(\mathbf{x})=f(\mathbf{x})$会方便，从而得到密度加权平均导数估计量[**density weighted average derivative (DWAD) estimator**]。

$$\hat{\boldsymbol{\delta}}_{\mathrm{DWAD}}=-\frac{1}{N}\sum_{i=1}^{N}y_i\hat{f}'(\mathbf{x}_i) \tag{9.41}$$

该式不再用 $\hat{f}(\mathbf{x}_i)$除。进而，得到 $\boldsymbol{\beta}$的$\sqrt{N}$一致的且渐近正态的估计值，只是至多差一个标度而已。例如，如果 $\boldsymbol{\beta}$的第一个分量被正规化为 1，那么对于 $j>1$，有 $\hat{\beta}_1=1$ 且 $\hat{\beta}_j=\hat{\delta}_j/\hat{\delta}_1$。

这些方法都要求连续回归元，从而导数存在。霍罗威茨和哈德尔（Horowitz and Härdle，1996）阐述过对离散回归元的推广。

半参数最小二乘法

一种可供选择的估计量是由市村（Ichimura，1993）提出的单指标模型。若以假定 $g(\cdot)$ 是已知的开始，在此情况下，$\hat{\boldsymbol{\beta}}$ 的 WLS 估计量是对

$$S_N(\boldsymbol{\beta})=\frac{1}{N}\sum_{i=1}^{N}w_i(x)(y_i-g(\mathbf{x}_i'\boldsymbol{\beta}))^2$$

求极小值。对于未知 $g(\cdot)$ 来说，市村提出用非参数估计值 $\hat{g}(\mathbf{x}_i'\boldsymbol{\beta})$ 代替 $g(\mathbf{x}_i'\boldsymbol{\beta})$，得出加权半参数最小二乘法[**weighted semiparametric least-squares (WSLS) estimator**]估计量 $\hat{\boldsymbol{\beta}}_{\text{WSLS}}$，它对

$$Q_N(\boldsymbol{\beta})=\frac{1}{N}\sum_{i=1}^{N}\pi(x_i)w_i(x)(y_i-\hat{g}(\mathbf{x}_i'\boldsymbol{\beta}))^2$$

求极小值，其中，$\pi(x_i)$ 表示修饰函数，当纯量 $\mathbf{x}_i'\boldsymbol{\beta}$ 的核回归估计值很小时，$\pi(x_i)$ 就省略了一些观测值，而 $\hat{g}(\mathbf{x}_i'\boldsymbol{\beta})$ 表示来自 y_i 对 $\mathbf{x}_i'\boldsymbol{\beta}$ 的回归的去掉一个的核估计量。这是 $\boldsymbol{\beta}$ 的 $\sqrt{N}$—一致且渐近正态估计，只是至多差一个标度，它通常比 DWAD 估计量更为有效。对于异方差数据来说，最有效的估计量是与可行的 GLS 相类似的估计量，它使用了估计加权函数 $\hat{w}_i(x)=1/\hat{\sigma}_i^2$，其中，$\hat{\sigma}_i^2$ 表示由 9.7.6 节中式(9.43)给出的核估计，这里，$\hat{u}_i=y_i-\hat{g}(\mathbf{x}_i'\hat{\boldsymbol{\beta}})$，而 $\hat{\boldsymbol{\beta}}$ 可由满足 $w_i(x)=1$ 的 $Q_N(\boldsymbol{\beta})$ 的最初极小化而获得。

通过迭代法，可计算 WSLS 估计量。若以初始估计量 $\hat{\boldsymbol{\beta}}^{(1)}$ 开始，譬如 DWAD 估计量的第一个分量被正规化为 1。由 $\hat{g}(\mathbf{x}_i'\hat{\boldsymbol{\beta}}^{(1)})$ 的核估计进而 $Q_N(\hat{\boldsymbol{\beta}}^{(1)})$ 知，为了获得梯度 $g_N(\hat{\boldsymbol{\beta}}^{(1)})=\partial Q_N(\boldsymbol{\beta})/\partial\boldsymbol{\beta}|_{\hat{\boldsymbol{\beta}}^{(1)}}$，要扰动 $\hat{\boldsymbol{\beta}}^{(1)}$，从而得到更新 $\hat{\boldsymbol{\beta}}^{(2)}=\hat{\boldsymbol{\beta}}^{(1)}+\mathbf{A}_N g_N(\hat{\boldsymbol{\beta}}^{(1)})$，等等。特别地，当 $Q_N(\boldsymbol{\beta})$ 是非凸且多峰的时候，这个估计量在计算上与 DWAD 估计量相比，就显得相当困难。

9.7.5 *广义加法模型*

广义加法模型(**generalized additive models**)设定 $\mathrm{E}[y|\mathbf{x}]=g_1(x_1)+\cdots+g_k(x_k)$，即完全非参数模型 $\mathrm{E}[y|\mathbf{x}]=g(x_1,\cdots,g_k)$ 的一种特殊化。这种特殊化导致被估计的子函数 $\hat{g}_j(x_j)$ 以一维非参数回归的速率收敛，而不是以 k 维非参数回归的较低速率收敛。

估计这类模型已有完善的方法[参见黑斯蒂和蒂伯沙拉尼（Hastie and Tibsharani，1990）]。在一些统计软件包中诸如 S-Plus，这是自动实施的。被估计的子函数 $\hat{g}_j(x_j)$ 对 x_j 的曲线勾画出 x_j 关于 $\mathrm{E}[y|\mathbf{x}]$ 的边际效应，因而加法模型能提供用于探索性数据分析的有益工具。这种模型在微观经济计量学中较少使用，部分原因在于，诸如删失、截取以及离散结果的应用常常会导致单指标模型与偏线性模型。

9.7.6 *异方差线性模型*

异方差线性模型(**heteroskedastic linear model**)设定：

$$E[y|\mathbf{x}]=\mathbf{x}'\boldsymbol{\beta}$$
$$V[y|\mathbf{x}]=\sigma^2(\mathbf{x})$$

其中，方差函数 $\sigma^2(\cdot)$ 是未设定的。

在现代微观经济计量学中，误差为异方差的假设是标准的横截面数据的假设。通过利用 OLS，并且使用 OLS 估计量方差矩阵的艾克—怀特(Eicker - White)异方差一致的估计，人们能获得 $\boldsymbol{\beta}$ 的一致却无效的估计。克拉格(Cragg，1983)以及雨宫(Amemiya，1983)曾经提出比 OLS 更为有效的工具变量估计量，但它仍不是完全有效的。虽然可行 GLS 已经提供完全有效的二阶矩估计量，可是它并不吸引人，因为它需要对 $\sigma^2(\mathbf{x})$ 的函数形式进行设定，譬如 $\sigma^2(\mathbf{x})=\exp(\mathbf{x}'\boldsymbol{\gamma})$。

鲁宾逊(Robinson，1987)提出利用 $\sigma_i^2=\sigma^2(\mathbf{x}_i)$ 的非参数估计量的 FGLS 的一种变形。于是有：

$$\hat{\boldsymbol{\beta}}_{\mathrm{HLM}}=\left(\sum_{i=1}^{N}\hat{\sigma}_i^{-2}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^{N}\hat{\sigma}_i^{-2}\mathbf{x}_i y_i\right) \tag{9.42}$$

这里，鲁宾逊(Robinson，1987)使用具有均匀权数的 σ_i^2 的 k - NN 估计量，所以：

$$\hat{\sigma}_i^2=\frac{1}{k}\sum_{j=1}^{N}\mathbf{1}(\mathbf{x}_j\in N_k(\mathbf{x}_i))\hat{u}_j^2 \tag{9.43}$$

其中，$\hat{u}_i=y_i-\mathbf{x}_i'\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ 表示来自 y_i 对 x_i 的第一阶段 OLS 回归的残差，而 $N_k(\mathbf{x}_i)$ 表示以欧几里德范数接近于 $\mathbf{x}_i$ 的 $\mathbf{x}_j$ 的 k 个观测值集合。于是，一旦假定 u_i 服从 iid $[0, \sigma^2(\mathbf{x}_i)]$，则：

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{\mathrm{HLM}}-\boldsymbol{\beta})\xrightarrow{d}\mathcal{N}\left[\mathbf{0},\ \left(\operatorname{plim}\frac{1}{N}\sum_{i=1}^{N}\sigma^{-2}(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\right]$$[1]

该估计量是适应的，因为它达到高斯—马尔可夫界，当 σ_i^2 已知时，它与 GLS 估计量一样有效。通过 $(N^{-1}\sum_i\hat{\sigma}_i^{-2}\mathbf{x}_i\mathbf{x}_i')^{-1}$，可一致估计出其方差矩阵。

原则上，还可使用 $\sigma^2(\mathbf{x}_i)$ 的其他一些非参数估计量，可是卡罗尔(Carroll，1982)和其他一些研究者最初提出使用 σ_i^2 的核估计量，而且发现，对有效性证明仅仅在对 $\mathbf{x}_i$ 有非常强的约束假设下才是可能的。鲁宾逊方法可被推广到具有非线性均值函数的模型。

9.7.7 半参数 MLE

假定 y_i 服从 iid，并具有设定密度 $f(y_i|\mathbf{x}_i,\boldsymbol{\beta})$。一般地讲，对密度错误设定会产生非一致参数估计值。加伦特和尼奇卡(Gallant and Nychka，1987)曾经提出，通过关于密度 $f(y|\mathbf{x},\boldsymbol{\beta})$ 的幂级数展开式来逼近未知真实密度。为了确保正密度，他们实际上使用 $f(y|\mathbf{x},\boldsymbol{\beta})$ 的平方幂级数展开(**squared power-series expansion**)，得到：

---

[1] 原著中该式出现符号上的错误，译者在此已经更正。——译者注

$$h_p(y|\mathbf{x},\boldsymbol{\beta},\boldsymbol{\alpha})=\frac{(p(y|\boldsymbol{\alpha}))^2 f(y|\mathbf{x},\boldsymbol{\beta})}{\int(p(z|\boldsymbol{\alpha}))^2 f(y|z,\boldsymbol{\beta})dz} \tag{9.44}$$

其中，$p(y|\boldsymbol{\alpha})$表示 y 的第 p 阶多项式，$\boldsymbol{\alpha}$表示多项式的系数向量，而除以分母则是要确保概率积分或者求和为 1。$\boldsymbol{\beta}$与 $\boldsymbol{\alpha}$ 的估计量是对对数似然$\sum_{i=1}^N \ln h_p(y_i|\mathbf{x},\boldsymbol{\beta},\boldsymbol{\alpha})$求极大值。该方法立刻被推广到多元变量 y_i 上。此估计量被称为半非参数极大似然估计量(**seminonparametric maximum likelihood estimator**)，因为它是一个非参数估计量，这可利用与极大似然估计量相同的方式加以估计。加伦特和尼奇卡(Gallant and Nychka，1987)已经证明，在相当一般的条件下，如果多项式的阶数 p 随着样本量 N 以适当速率递增，那么估计量会得到密度的一致估计值。

为了获得任何特殊数据的灵活分布，这一结果提供了利用式(9.44)的一种坚实基础。若关于基准密度 $f(y|\mathbf{x},\boldsymbol{\beta})$的多项式序列 $p(y|\boldsymbol{\alpha})$是正交的或标准正交多项式序列(参见 12.3.1 节)，则该方法特别简单，从而分母中的正规化因子能直接构造。利用信息准则选取多项式的阶，就惩罚模型的测量而言，其复杂性大于实际应用时的 AIC。当人们忽略对多项式阶的数据相依选取，同时假定得到的密度 $h_p(y|\mathbf{x},\boldsymbol{\beta},\boldsymbol{\alpha})$被正确设定，通常的 ML 统计推断就是可行的。对于计数回归来说，该方法的例子由卡梅伦和约翰森(Cameron and Johansson，1997)给出。

9.7.8 半参数有效界

半参数有效界(**semiparametric efficiency bounds**)是将有效性譬如克莱默—劳或者高斯—马尔可夫定理推广到数据生成过程(dgp)具有非参数成分的情况。最佳半参数方法就达到了这个有效界。

我们用 $\boldsymbol{\beta}$表示想要估计的参数，可能包括方差成分譬如 σ^2，而用 $\boldsymbol{\eta}$ 表示冗余参数。为了简单起见，我们考察具有非参数成分的极大似然估计。

我们以完全参数情况开始。MLE($\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\eta}}$)对 $\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\eta})=\ln L(\boldsymbol{\beta},\boldsymbol{\eta})$求极大值。设 $\boldsymbol{\theta}=(\boldsymbol{\beta},\boldsymbol{\eta})$，并设 $\mathcal{I}_{\theta\theta}$表示式(5.43)定义的信息矩阵。于是，$\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})\xrightarrow{d}\mathcal{N}[\mathbf{0},\mathcal{I}_{\theta\theta}^{-1}]$。对于$\sqrt{N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$来说，当 $\boldsymbol{\eta}$ 已知时，$\mathcal{I}_{\theta\theta}$的分块反演导致

$$\mathrm{V}^*=(\mathcal{I}_{\beta\beta}-\mathcal{I}_{\beta\eta}\mathcal{I}_{\eta\eta}^{-1}\mathcal{I}_{\eta\beta})^{-1} \tag{9.45}$$

作为估计 $\boldsymbol{\beta}$的有效界。当 $\boldsymbol{\eta}$ 未知时，存在有效性损失，除非信息矩阵是分块对角的，因此，$\mathcal{I}_{\beta\eta}=\mathbf{0}$ 且方差简化成 $\mathcal{I}_{\beta\beta}^{-1}$。

现在，考虑对非参数情况的推广。假定我们具有参数子模型，比如说 $\mathcal{L}_0(\boldsymbol{\beta})$，这只涉及 $\boldsymbol{\beta}$。考察对某一个 $\boldsymbol{\eta}$ 值嵌入 $\mathcal{L}_0(\boldsymbol{\beta})$的全部可能参数模型 $\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\eta})$的族。在所有可能参数模型 $\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\eta})$上，半参数有效界是式(9.45)给出的 V^* 的最大值，可是这个界难以求出。

考察：

$$\bar{\mathbf{s}}_\beta=\mathbf{s}_\beta-\mathrm{E}[\mathbf{s}_\beta|\mathbf{s}_\eta]$$

进行简化是可能的，其中，$\mathbf{s}_\theta$ 表示得分$\partial\mathcal{L}/\partial\boldsymbol{\theta}$，而 $\bar{\mathbf{s}}_\beta$ 表示剔除 $\boldsymbol{\eta}$ 之后 $\boldsymbol{\beta}$的得分。对于有限维的 $\boldsymbol{\eta}$ 来说，可以证明，$\mathrm{E}[N^{-1}\bar{\mathbf{s}}_\beta\bar{\mathbf{s}}_\beta']=\mathrm{V}^*$。不过，这里的 $\boldsymbol{\eta}$ 是无限维的。

假定数据为 iid 的，并设 $\mathbf{s}_{\theta i}$ 表示和式中导致得分 $\mathbf{s}_{\theta}$ 的第 i 个分量。贝根等人(Begun et al.，1983)把切集(tangent set)定义成 $\mathbf{s}_{\eta i}$ 的所有线性组合的集合。当切集是线性的且是闭的，(9.45)式 V^* 的最大值等于：

$$\mathbf{\Omega}=(\operatorname{plim} N^{-1}\tilde{\mathbf{s}}_{\beta}\tilde{\mathbf{s}}_{\beta}')^{-1}=(\mathrm{E}[\tilde{\mathbf{s}}_{\beta i}\tilde{\mathbf{s}}_{\beta i}'])^{-1}$$

于是，矩阵 $\mathbf{\Omega}$ 是半参数有效界。

在应用时，人们首先求出 $\mathbf{s}_{\eta}=\sum_i \mathbf{s}_{\eta i}$。然后求 $\mathrm{E}[\mathbf{s}_{\beta i}|\mathbf{s}_{\eta i}]$，这就需要譬如误差对称性的假设，而这些假设是对所要考察的半参数模型类上施加的约束。这就得出 $\tilde{\mathbf{s}}_{\beta i}$，从而得到 $\mathbf{\Omega}$。对于更详细内容及应用，参见纽韦(Newey，1990)、帕甘和乌拉(Pagan and Ullah，1999)以及塞韦林尼和特里帕蒂(Severini and Tripathi，2001)。

9.8 核估计量均值与方差推导

非参数估计需要在光滑性(方差)与偏倚(均值)之间进行权衡。这里，我们推导核密度与核回归估计量的均值以及方差。推导将沿着李明宰(M. J. Lee，1996)的那些线索而展开。

9.8.1 *核密度估计量均值与方差*

由于 x_i 是 iid 的，故求和式中的每一项都具有相同期望值，并且：

$$\begin{aligned}\mathrm{E}[\hat{f}(x_0)]&=\mathrm{E}\left[\frac{1}{h}K\left(\frac{x-x_0}{h}\right)\right]\\&=\int\frac{1}{h}K\left(\frac{x-x_0}{h}\right)f(x)dx\end{aligned}$$

通过对 $z=(x-x_0)/h$ 进行变量变换，因此 $x=x_0+hz$，从而 $dx/dz=h$，得到：

$$\mathrm{E}[\hat{f}(x_0)]=\int K(z)f(x_0+hz)dz$$

由 $f(x_0+hz)$ 在 $f(x_0)$ 处的二阶泰勒级数展开式，得到：

$$\begin{aligned}\mathrm{E}\,[\hat{f}(x_0)]&=\int K(z)\left\{f(x_0)+f'(x_0)hz+\frac{1}{2}f''(x_0)(hz)^2\right\}dz\\&=f(x_0)\int K(z)dz+hf'(x_0)\int zK(z)dz+\frac{1}{2}h^2f''(x_0)\int z^2K(z)dz\end{aligned}$$

由于核 $K(z)$ 积分为 1，故上式简化为：

$$\mathrm{E}[\hat{f}(x_0)]-f(x_0)=hf'(x_0)\int zK(z)dz+\frac{1}{2}h^2f''(x_0)\int z^2K(z)dz$$

如果除了核满足 $\int zK(z)dz=0$ 外，还具有 9.3.3 节条件(ii) 中所做的假定，同时 f 的二阶导数是有界的，那么右边第一项就会消失，从而得到，$\mathrm{E}[\hat{f}(x_0)]-f(x_0)=b(x_0)$，其中，$b(x_0)$ 已在式(9.4) 中定义。

为了获得 $\hat{f}(x_0)$的方差，以下面注意到的内容开始：若 y_i 是 iid 的，则 $\mathrm{V}[\bar{y}]=$

$N^{-1}\mathrm{V}[y]=N^{-1}\mathrm{E}[y^2]-N^{-1}(E[y^2])^2$。因而：

$$\mathrm{V}[\hat{f}(x_0)]=\frac{1}{N}\mathrm{E}\left[\left(\frac{1}{h}K\left(\frac{x-x_0}{h}\right)\right)^2\right]-\frac{1}{N}\left(\mathrm{E}\left[\frac{1}{h}K\left(\frac{x-x_0}{h}\right)\right]\right)^2$$

现在，通过变量变化与一阶泰勒级数展开：

$$\begin{aligned}\mathrm{E}\left[\left(\frac{1}{h}K\left(\frac{x-x_0}{h}\right)\right)^2\right]&=\int\frac{1}{h}K(z)^2\{f(x_0)+f'(x_0)hz\}dz\\&=\frac{1}{h}f(x_0)\int K(z)^2dz+f'(x_0)\int zK(z)^2dz\end{aligned}$$

由此可得：

$$\begin{aligned}\mathrm{V}[\hat{f}(x_0)]=&\frac{1}{Nh}f(x_0)\int K(z)^2dz+\frac{1}{N}f'(x)\int zK(z)^2dz\\&-\frac{1}{N}\left[f(x_0)+\frac{h^2}{2}f'(x_0)\left[\int z^2K(z)dz\right]\right]^2\end{aligned}$$

当 $h\to0$ 且 $N\to\infty$ 时，这由第一项来控制，从而得到式(9.5)。

9.8.2 核回归估计量分布

我们想要获得回归元 x_i 的分布，而 x_i 是 iid 的且具有密度 $f(x)$。由 9.5.1 节知，核估计量是一种加权平均 $\hat{m}(x_0)=\sum_i w_{i0,h}y_i$，其中，核权数 $w_{i0,h}$ 已由式(9.22)给出。由于权数之和为 1，所以有 $\hat{m}(x_0)-m(x_0)=\sum_i w_{i0,h}(y_i-m(x_0))$。将式(9.15)代入 y_i 中，并且如同核密度估计量一样用 $\sqrt{Nh}$ 去正规化，得出：

$$\sqrt{Nh}(\hat{m}(x_0)-m(x_0))=\sqrt{Nh}\sum_{i=1}^{N}w_{i0,h}(m(x_i)-m(x_0)+\varepsilon_i)\quad(9.46)$$

一种获得式(9.46)的极限分布方法是，取 $m(x_i)$ 在 x_0 附近的二阶泰勒级数展开。这种方法并不总行得通，因为正规化的缘故，其权数之和为 1，权数 $w_{i0,h}$ 变得十分复杂[参见式(9.22)]。

不过，我们遵循比尔恩斯(Bierens，1987，第 106～108 页)思想，采用李明宰(Lee，1996，第 148～151 页)的方法。注意到，由于 $\hat{f}(x_0)=(Nh)^{-1}\sum_i K((x_i-x_0)/h)$，所以加权函数的分母是 x_0 的密度的核估计。于是，式(9.46)变为：

$$\sqrt{Nh}(\hat{m}(x_0)-m(x_0))=\frac{1}{\sqrt{Nh}}\sum_{i=1}^{N}K\left(\frac{x_i-x_0}{h}\right)(m(x_i)-m(x_0)+\varepsilon_i)/\hat{f}(x_0)\quad(9.47)$$

我们把变换定理(定理 A.12)用于式(9.47)，对于分母利用式(9.47)，为了得到分子的极限正态分布，需要下面几步推导：

$$\begin{aligned}&\frac{1}{\sqrt{Nh}}\sum_{i=1}^{N}K\left(\frac{x_i-x_0}{h}\right)(m(x_i)-m(x_0)+\varepsilon_i)\quad(9.48)\\&=\frac{1}{\sqrt{Nh}}\sum_{i=1}^{N}K\left(\frac{x_i-x_0}{h}\right)(m(x_i)-m(x_0))+\frac{1}{\sqrt{Nh}}\sum_{i=1}^{N}K\left(\frac{x_i-x_0}{h}\right)\varepsilon_i\end{aligned}$$

考察式(9.48)中的第一个和式;如果可应用大数极限定律,那么它将依概率收敛到均值:

$$\begin{aligned}
&\mathrm{E}\left[\frac{1}{\sqrt{Nh}}\sum_{i=1}^{N}K\left(\frac{x_i-x_0}{h}\right)(m(x_i)-m(x_0))\right] \qquad (9.49)\\
&=\frac{\sqrt{N}}{\sqrt{h}}\int K\left(\frac{x-x_0}{h}\right)(m(x)-m(x_0))f(x)dx\\
&=\sqrt{Nh}\int K(z)(m(x_0+hz)-m(x_0))f(x_0+hz)dz\\
&=\sqrt{Nh}\int K(z)\left(hzm'(x_0)+\frac{1}{2}h^2z^2m''(x_0)\right)(f(x_0)+hzf'(x_0))dz\\
&=\sqrt{Nh}\left\{\int K(z)h^2z^2m'(x_0)f'(x_0)dz+\int K(z)\ \frac{1}{2}h^2z^2m''(x_0)f(x_0)dz\right\}\\
&=\sqrt{Nh}h^2\left(m'(x_0)f'(x_0)+\frac{1}{2}m''(x_0)f(x_0)\right)\int z^2K(z)dz\\
&=\sqrt{Nh}f(x_0)b(x_0)
\end{aligned}$$

其中,$b(x_0)$已由式(9.23)定义。第一个等式使用 x_i 为 iid 的;第二个等式是对 z 做一个变量变换 $z=(x-x_0)/h$;第三个等式则对 $m(x_0+hz)$ 应用二阶泰勒展开,并对 $f(x_0+hz)$应用一阶泰勒级数展开;第四个等式成立,是因为给定控制其他项的两项时,把乘积展开成四项[例如,参见李明宰(Lee, 1996,第 150 页)]。

现在,考察式(9.48)中的第二个和式;显然,和式中具有零均值,而每一项的方差在省略下标后变成:

$$\begin{aligned}
\mathrm{V}\left[K\left(\frac{x-x_0}{h}\right)\varepsilon\right]&=\mathrm{E}\left[K^2\left(\frac{x-x_0}{h}\right)\varepsilon^2\right] \qquad (9.50)\\
&=\int K^2\left(\frac{x-x_0}{h}\right)\mathrm{V}[\varepsilon|x]f(x)dx\\
&=h\int K^2(z)\mathrm{V}[\varepsilon|x_0+hz]f(x_0+hz)dz\\
&=h\mathrm{V}[\varepsilon|x_0]f(x_0)\int K^2(z)dz
\end{aligned}$$

在第三行中,对 z 做一个变量变换 $z=(x-x_0)/h$,且 $dx=hdz$,同时令 $h\to0$ 来得到最后一行。利用中心极限定理,由此可得:

$$\frac{1}{\sqrt{Nh}}\sum_{i=1}^{N}K\left(\frac{x_i-x_0}{h}\right)\varepsilon_i\xrightarrow{d}\mathcal{N}\left[0,\ \mathrm{V}[\varepsilon|x_0]f(x_0)\int K^2(z)dz\right] \qquad (9.51)$$

当把式(9.49)与式(9.51)结合起来,我们得出在式(9.47)中定义的 $\sqrt{Nh}(\hat{m}(x_0)-m(x_0))$ 收敛于 $1/f(x_0)$ 倍的 $\mathcal{N}[\sqrt{Nh}f(x_0)b(x_0),\ \mathrm{V}[\varepsilon|x_0]\times f(x_0)\int K^2(z)dz]$。用 $f(x_0)$ 除均值,并用 $f(x_0)^2$ 除方差,得到式(9.24)给出的极限分布。

9.9 应用研究

适合于各种用途的回归软件日趋增多，这为单变量非参数密度估计与回归提供了足够的方法。程序语言 XPlore 强调非参数方法以及画图法；许多方法的详细内容，已在其网站上提供。

利用建立在核基础上的核密度估计，诸如高斯核或埃帕内尼科夫核，对非参数单变量密度进行估计就简单易行。容易计算的插值估计，为带宽提供了一个的有益起点，比如说该起点可能缩减一半或增大一倍，这要视其改进而定。

若不管带宽选取，则非参数单变量回归也是简单易行的。如果回归函数在端点处的相对无偏估计值是人们所期望的，那么局部线性回归或洛斯估计都比核回归要好。关于带宽的插值估计相当难以获得，却可以使用交叉验证（参见 9.5.3 节），以及盯住散点图和拟合直线。人们期望的光滑性程序随着应用而变化。对于非参数多变量回归来说，这种视力观察是难以做到的。

就半参数回归而言，其内容更加复杂。由于参数成分的典型估计包含对非参数成分的一种平均，所以半参数回归需要一些技巧，诸如对非参数成分进行修饰与光滑不足。为了这种目的，人们通常运用以诸如 Gauss、Matlab、Splus 或 XPlore 语言编写的特定程序。对于非参数估计成分来说，使用快速计算算法能节省相当多的计算量，例如重新分级与调整（**binning and updating**），参见范剑青和吉贝尔斯（Fan and Gijbels，1996）以及哈德尔和林顿（Härdle and Linton，1994）。

有时，所有方法都需要对带宽或窗口宽度加以设定。各种不同选取会导致有限样本出现不同的估计值，其差异相当大，正如本章中的一些图形所阐述的那样。与之相比，在完全参数框架下，不同研究者通过极大似然法估计同一模型都将得到一样的参数估计值。这种不确定性是对非参数方法的贬低，尽管希望是在半参数方法中至少影响模型的参数成分的效果或许是很小的。

9.10 文献注释

非参数估计在许多统计学教科书中得到了很好阐述，包括范剑青和吉贝尔斯（Fan and Gijbels，1996）。鲁珀特、万德和卡罗尔（Ruppert，Wand and Carroll，2003）曾经阐述许多半参数方法的应用。由哈德尔（Härdle，1990）、李明宰（M. J. Lee，1996）、霍罗维茨（Horowitz，1998b）、帕甘和乌拉（Pagan and Ullah，1999）以及亚特丘（Yatchew，2003）撰写的经济计量学教科书既涵盖非参数估计，又涵盖半参数估计。亚特丘（Yatchew，2003）的书则是面向应用经济计量家而撰写的。他强调偏线性模型与单指标模型，还有上述模型诸如置信区间计算的实际问题。

9.3 核密度估计的重要早期文献是，罗森布拉特（Rosenblatt，1956）和帕曾（Parzen，1962）。西尔弗曼（Silverman 1986）的书是非参数密度估计方面的经典书。

9.4 对非参数估计量的最优收敛速率进行更一般的研究由斯通（Stone，1980）给出。

9.5 核回归估计量是由纳达雷娅(Nadaraya, 1964)与沃森(Watson, 1964)提出的。核与最近邻回归的一个非常有用的且相对简单的综述则是由奥尔特曼(Altman, 1992)给出。统计学文献中,存在许多其他综述。哈德尔(Härdle, 1990, 第5章)曾经对带宽选取与置信区间提供了详细阐述。

9.6 关于半参数有效界,参见由纽韦(Newey, 1990b)撰写的一篇综述,而最新的论文则出自塞韦林尼和特里帕蒂(Severini and Tripathi, 2001)。早期的经济计量应用由张伯伦(Chamberlain, 1987)给出。

9.6 非参数局部回归的许多方法都包含在斯通(Stone, 1997)的论文中。关于序列估计量,参见安德鲁斯(Andrews, 1991)与纽韦(Newey, 1997)。

9.7 经济计量学文献关注于半参数回归。综述性论文包括由鲍威尔(Powell, 1994)、鲁宾逊(Robinson, 1988b)所撰写的那些论文,而在更为导论性水平上的是亚特丘的书(Yatchew, 1998)。本书的其他一些地方譬如14.7节、15.11节、16.9节、20.5节以及23.8节都曾给出另外的一些参考文献。贝尔马、梅伦伯格以及范·索斯特(Bellmare, Melenberg and Van Soest, 2002)做出了一项应用研究,他们阐述过几种半参数方法。

习 题

9-1 假定我们利用均匀核(参见表9.1),满足 $h=1$ 且样本量 $N=100$,可获得核密度估计。假定实际上数据 $x\sim\mathcal{N}[0, 1]$。

(a) 利用式(9.4),计算核密度估计在 $x_0=1$ 的偏倚。

(b) 此偏倚相对于真实值 $\phi(1)$ 而言,显得大吗? 其中,$\phi(\cdot)$表示标准正态 pdf。

(c) 利用式(9.5),计算核密度估计在 $x_0=1$ 的方差。

(d) 对于方差和偏倚平方,哪一个对 MSE 在 $x_0=1$ 的值做出更大贡献?

(e) 利用9.3.7节的结果,给出建立在核密度估计值 $\hat{f}(1)$基础上的密度在 $x_0=1$ 处的95%置信区间。

(f) 对本题而言,由式(9.10),什么值是最优带宽 h^*?

9-2 假定我们利用均匀核(参见表9.1),满足 $h=1$ 且样本量 $N=100$,可获得核密度估计。假定数据实际上 $x\sim\mathcal{N}[0, 1]$,并且条件均值函数是 $m(x)=x^2$。

(a) 利用式(9.23),计算核回归估计在 $x_0=1$ 的偏倚。

(b) 此偏倚相对于真实性 $m(1)=1$ 而言显得大吗?

(c) 利用式(9.24),计算核回归估计在 $x_0=1$ 的方差。

(d) 对于方差和偏倚平方,哪一个对 MSE 在 $x_0=1$ 处的值做出贡献更大?

(e) 利用9.5.4节的结果,给出建立在核回归估计值 $m(1)$基础上 $\mathrm{E}[y|x_0=1]$ 的95%置信区间。

9-3 假定这一问题可使用非参数密度估计方案。运用4.6.4节关于健康消费数据。利用具有高斯核(如果有的)的核密度估计。

(a) 通过目测观察,并用试错法选取合适带宽,求关于健康消费的核密度估计,叙述带宽的选取。

(b) 通过目测观察,并用试错法选取合适带宽,求健康消费的自然对数的核密度估计。叙述带宽的选取。

(c) 把你在(b)部分的解答与适当的直方图对比。

(d) 如果可能,将拟合正态密度叠放到与来自(b)部分核密度估计的同一图形上。健康消费看起来会是对数正态分布吗?

9-4 假定这个问题可使用核回归方案或其他非参数光滑子。运用4.6.4节关于健康消费(y)的自然对数与总消费(x)的自然对数数据的完整样本。

(a) 一旦通过目测观察,并用试错法选取良好的带宽,求关于健康消费的核回归估计[1]。叙述带宽的选取。

(b) 给定(a)部分,健康看起来会是正态商品吗?

(c) 给定(a)部分,健康看起来会是奢侈商品吗?

(d) 把你的非参数估计与来自线性回归及二次回归的预测进行比较。

---

[1] 原著中这里为"核回归密度估计",应为"核回归估计",已改。——译者注

10 数值最优化

10.1 引 论

第 5 章和第 6 章已经阐述，把估计量的一致性及渐近分布定义成最优化问题解的有关理论结果。一个更为实际的问题是，如何获得数值最优解，也就是说，当估计量不存在显式公式时如何计算参数估计量，这将构成本章主题。

对于应用研究者来说，标准的非线性模型比如 logit、Tobit、比例风险以及泊松模型的估计，看起来似乎与 OLS 模型的估计并没有什么差异。利用统计软件可获得估计值，并报告系数、标准误差、t 统计量以及 p 值。一般地讲，只有因 OLS 失效，譬如出现多重共线性或不正确的数据输入，才会引发需要计算的问题。

对那些缺少标准的非线性模型包括标准模型的稍微变形进行估计，都需要编写程序。这在标准的统计软件中或许是可行的。否则，就要使用编程语言。尤其是在后一种情况下，必须具备最优化方法的知识。

10.2 节对最优化给出一般性研究。各种各样的迭代法包括牛顿—拉夫森(Newton - Raphson)、高斯—牛顿(Gauss - Newton)梯度法将在 10.3 节加以阐述。一些实际问题，像某些普遍易犯的错误，则在 10.4 节讨论。当用最优化方法不能得出参数估计值时，这些问题就显得尤其有意义。

10.2 一般性研究

微观经济计量分析时常建立在估计量 $\hat{\boldsymbol{\theta}}$ 基础上，该估计量针对随机目标函数 $Q_N(\boldsymbol{\theta})$求极大值，$\hat{\boldsymbol{\theta}}$ 通常是一阶条件$\partial Q_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}=\mathbf{0}$ 的解。求极小值问题可通过用 -1 乘以目标函数而改为求极大值。在非线性应用即 q 个方程关于 q 个未知 $\boldsymbol{\theta}$ 的非线性方程组中，一阶条件通常不存在显式解。

通常格点搜索程序行不通，而迭代法即通常的梯度法却行之有效。

10.2.1 *格点搜索*

就格点搜索法(**grid starch methods**)而言，程序(**procedure**)为沿着格点选取许多 $\boldsymbol{\theta}$ 的不同值，对这些 $\boldsymbol{\theta}$ 值中的每一个都要进行计算 $Q_N(\boldsymbol{\theta})$，并选择估计量 $\hat{\boldsymbol{\theta}}$ 作为

使 $Q_N(\boldsymbol{\theta})$(局部或全局依赖于应用问题)成为最大值的那一个值。

如果能选取足够精细的网格,那么这种方法总会起作用。然而,若没有进一步限制,选取足够精细的网格,通常是不切实际的。例如,当有 10 个参数要估计时,网格对于每一个参数都恰好在 10 个点即非常稀疏网格上进行计算,这将会有 10^{10} 或 100 亿个计算值。

不过,格点搜索法在下面一些应用中却是有益的,即格点搜索只需在参数的一个子集上加以搜索计算。为了在使用迭代法时,人们不必担忧出现多重最大值问题,格点搜索要通过检查响应面[1](**response surface**)来验证这一点。例如,许多时间序列软件对具有 AR(1)误差的回归模型的纯量 AR(1)系数就是这样做的。第二个例子是,对嵌套 logit 模型(参见 15.6 节)的纯量相容系数(**inclusive parameter**)实施格点搜索。当然,若其他什么方法都不起作用,则必须用格点搜索法。

10.2.2 *迭代法*

实际上,所有微观经济计量在应用时反而都使用迭代法(**iterative methods**)。这些迭代法利用特定规则,不断更新当前 $\boldsymbol{\theta}$ 估计值。已知第 s 次估计值 $\hat{\boldsymbol{\theta}}_s$,迭代法提供可产生新估计值 $\hat{\boldsymbol{\theta}}_{s+1}$ 的一个规则,其中,$\hat{\boldsymbol{\theta}}_s$ 表示第 s 次估计值,而不是 $\hat{\boldsymbol{\theta}}$ 的第 s 个成分。原则上讲,新的估计值会向着最大值运动,因而有 $Q_N(\hat{\boldsymbol{\theta}}_{s+1})>Q_N(\hat{\boldsymbol{\theta}}_s)$,但是通常这一点无法得到保障。此外,梯度估计(值)或许找到局部最大值,但不一定是全局最大值。

10.2.3 *梯度法*

大多数的迭代法都是梯度法(**gradient methods**),即在梯度所确定的方向上对 $\hat{\boldsymbol{\theta}}_s$ 加以变动。一个校正公式是梯度

$$\hat{\boldsymbol{\theta}}_{s+1}=\hat{\boldsymbol{\theta}}_s+\mathbf{A}_s\mathbf{g}_s, \quad s=1,\cdots,s \tag{10.1}$$

的矩阵加权平均,其中,$\mathbf{A}_s$ 表示 $q\times q$ 阶矩阵,它依赖于 $\hat{\boldsymbol{\theta}}_s$,而:

$$\mathbf{g}_s=\left.\frac{\partial Q_N(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|_{\hat{\boldsymbol{\theta}}_s} \tag{10.2}$$

表示 $q\times 1$ 维梯度向量(**gradient vector**)在 $\hat{\boldsymbol{\theta}}_s$ 处的计算值。各种不同梯度法运用不同的矩阵 $\mathbf{A}_s$,其详细情况在 10.3 节阐述。一个重要的例子是牛顿—拉夫森方法,该方法设 $\mathbf{A}_s=-\mathbf{H}_s^{-1}$,其中,$\mathbf{H}_s$ 表示海赛矩阵,稍后将在式(10.6)中加以定义。注意到,本章的 $\mathbf{A}$ 与 $\mathbf{g}$ 表示数量,这有别于其他章节的符号内容。这里,$\mathbf{A}$ 不是估计量极限分布中所出现的矩阵,而 $\mathbf{g}$ 不是非线性回归模型中 $\mathbf{y}$ 的条件均值。

原则上讲,矩阵 $\mathbf{A}_s$ 对最大值而言是正定的(**positive definite**)(或对最小值而言是负定的),从而可能有 $Q_N(\hat{\boldsymbol{\theta}}_{s+1})>Q_N(\hat{\boldsymbol{\theta}}_s)$。这由一阶泰勒级数展开式 $Q_N(\hat{\boldsymbol{\theta}}_{s+1})=Q_N(\hat{\boldsymbol{\theta}}_s)+\mathbf{g}_s'(\hat{\boldsymbol{\theta}}_{s+1}-\hat{\boldsymbol{\theta}}_s)+R$ 可得,其中,R 表示余项。所以,一旦代入更新公式

[1] 又称为反应面。——译者注

(10.1)中，得到：

$$Q_N(\hat{\boldsymbol{\theta}}_{s+1})-Q_N(\hat{\boldsymbol{\theta}}_s)=\mathbf{g}_s'\mathbf{A}_s\mathbf{g}_s+R$$

如果 $\mathbf{A}_s$ 是正定的且余项R 充分小，上式就大于 0，因为对正定方阵 $\mathbf{A}$ 来说，对于所有列向量 $\mathbf{x}\neq\mathbf{0}$，二次式 $\mathbf{x}'\mathbf{A}\mathbf{x}>0$。太小的 $\mathbf{A}_s$ 值会使迭代程序太慢；不过，即使 $\mathbf{A}_s$ 是正定的，太大的 $\mathbf{A}_s$ 值会导致超过适当限度，因为就很大变动而言，不能忽略余项。

对梯度法的一种普遍修正是，添加步长调整(**step-size adjustment**)来防止可能超过适当限度或未达到适当限度，因此：

$$\hat{\boldsymbol{\theta}}_{s+1}=\hat{\boldsymbol{\theta}}_s+\hat{\lambda}_s\mathbf{A}_s\mathbf{g}_s \tag{10.3}$$

其中，步长 $\hat{\lambda}_s$ 表示选取使得 $Q_N(\hat{\boldsymbol{\theta}}_{s+1})$ 达到最大值的那个纯量。在第 s 次上，首先计算 $\mathbf{A}_s\mathbf{g}_s$，这会涉及相当的计算量。然后，计算 $Q_N(\hat{\boldsymbol{\theta}}_s)$，对于 λ 的取值范围来说(称为线搜索)，$\hat{\boldsymbol{\theta}}=\hat{\boldsymbol{\theta}}_s+\lambda\mathbf{A}_s\mathbf{g}_s$，并且像 λ 那样选取 $\hat{\lambda}_s$ 使得 $Q_N(\hat{\boldsymbol{\theta}})$ 最大化。因为梯度与 $\mathbf{A}_s$ 均不用沿着线搜索重新计算，故可节省相当多的计算量。

当矩阵 $\mathbf{A}_s$ 被定义成矩阵 $\mathbf{B}_s$ 的逆时，因此，比如说，$\mathbf{A}_s=\mathbf{B}_s^{-1}$，有时要做出第二次修正。于是，如果 $\mathbf{B}_s$ 接近于常值的奇异矩阵。比如说 $\mathbf{C}$，就要加上或减去 $\mathbf{C}$ 以使其逆存在，因而 $\mathbf{A}_s=(\mathbf{B}_s+\mathbf{C})^{-1}$。当 $\mathbf{A}_s$ 不是正定的，就要做类似的调整。对 $\mathbf{A}_s$ 的更进一步讨论，将在 10.3 节给出。

梯度法最有可能收敛到最靠近初始值的那个局部最大值。假如目标函数有多重局部最优值，则一系列的初始值将被用于增加寻找全局最大值的机会。

10.2.4 *梯度法例子*

考察当唯一的回归元是截距时，指数回归模型的 NLS 估计量的计算。于是，$\mathrm{E}[y]=e^{\beta}$，并经过一些代数运算，可得到梯度 $g=N^{-1}\sum_i(y_i-e^{\beta})e^{\beta}=(\bar{y}-e^{\beta})e^{\beta}$。假定在式(10.1)中，我们使用 $\mathbf{A}_s=e^{-2\hat{\beta}_s}$，这对应于稍后 10.3.2 节将阐述的牛顿—拉夫森算法的得分变形方法。迭代法简化成 $\hat{\beta}_{s+1}=\hat{\beta}_s+(\bar{y}-e^{\hat{\beta}_s})/e^{\hat{\beta}_s}$。

举一个执行这种算法的例子，假定 $\bar{y}=2$ 且初始值是 $\hat{\beta}_1=0$ 。这就得到表 10.1 所列的一些迭代。该例子非常迅速地收敛到 NLS 估计值，对这个简单例子而言，用解析方法能得到，$\hat{\beta}=\ln\bar{y}=\ln 2=0.693\,147$。目标函数自始至终地增大，它是对含有全局凹目标函数使用 NR 算法的结果。注意到，在第一次迭代即从 $\hat{\beta}_1=0.0$ 到 $\hat{\beta}_2=1.0$ 时，出现超过适当限度，大于 $\hat{\beta}=0.693$。

表 10.1 梯度法结果

次数	估计值	梯度	目标函数
s	$\hat{\beta}_s$	g_s	$Q_N(\hat{\beta}_s)=-\frac{1}{2N}\sum_i(y_i-e^{\beta})^2$
1	0.000 000	1.000 000	$1.500\,000-\sum_i y_i^2/2N$
2	1.000 000	−1.952 492	$1.742\,036-\sum_i y_i^2/2N$
3	0.735 758	−0.181 711	$1.996\,210-\sum_i y_i^2/2N$
4	0.694 042	−0.003 585	$1.999\,998-\sum_i y_i^2/2N$
5	0.693 147	−0.000 002	$2.000\,000-\sum_i y_i^2/2N$

当使用 NR 算法且目标函数是全局凹的时候，通常会出现快速收敛。实际应用时的一个挑战是，非标准非线性模型经常具有不是全局凹的目标函数。

10.2.5 *矩方法与 GMM 估计量*

对于 m 估计量来说，$Q_N(\boldsymbol{\theta})=N^{-1}\sum_i q_i(\boldsymbol{\theta})$，并且梯度 $\mathbf{g}(\boldsymbol{\theta})=N^{-1}\sum_i \partial q_i(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$。

对于广义矩方法估计量来说，$Q_N(\boldsymbol{\theta})$是一个二次型（参见 6.3.2 节），而梯度为更复杂的形式：

$$\mathbf{g}(\boldsymbol{\theta})=\left[N^{-1}\sum_i \partial\mathbf{h}_i(\boldsymbol{\theta})'/\partial\boldsymbol{\theta}\right]\times\mathbf{W}_N\times\left[N^{-1}\sum_i \mathbf{h}_i(\boldsymbol{\theta})\right]$$

于是，不能再使用某些梯度法，因为它们只是对平均起作用。10.3 节给出的一些方法，还是能被人们使用的，包括牛顿—拉夫森、最速下降法、DFP、BFG 以及模拟退火法。

矩阵法与估计方程估计量被定义成方程组的解，但它们类似于广义矩方法，能变换成数值最优化问题。求解 q 个方程 $N^{-1}\sum_i\mathbf{h}_i(\boldsymbol{\theta})=\mathbf{0}$ 的估计量，能通过对 $Q_N(\boldsymbol{\theta})=[N^{-1}\sum_i\mathbf{h}_i(\boldsymbol{\theta})]'[N^{-1}\sum_i\mathbf{h}_i(\boldsymbol{\theta})]$ 求最小值而得到。

10.2.6 *收敛准则*

迭代过程要不断进行，一直到不存在变化为止。当下面所有情形发生时，原则上程序应停止：(1)在目标函数 $Q_N(\hat{\boldsymbol{\theta}}_s)$中出现很小的相对变化；(2)相对于海赛矩阵来说，出现很小的梯度向量 $\mathbf{g}_s$ 的变化；(3)参数估计值 $\hat{\boldsymbol{\theta}}_s$ 中出现很小的相对变化。统计软件对这三种变化典型地选取默认的极限值，称之为收敛准则(**convergence criteria**)。这些值经常由使用者来变动。保守值取为 10^{-6}。

此外，通常存在试图达到最大的迭代次数(**maximum number of iterations**)。当达到这个最大值时，该估计量典型地被报告出来。可是，除非达到收敛，否则不应使用该估计值。

若达到收敛，则获得局部最大值。然而，除非目标函数是全局凹的，否则不能确保获得全局最大值。

10.2.7 *初始值*

如果最初的初始值(**starting value**)$\hat{\boldsymbol{\theta}}_1$ 接近于 $\hat{\boldsymbol{\theta}}$，那么迭代次数在很大程度上会得到减少。很明显，一致参数估计量是作为初始值的良好估计量。一个不好的初始值选取能导致迭代法失败。特别地，对于某些估计量与梯度法来说，当初始值是 $\hat{\boldsymbol{\theta}}_1=\mathbf{0}$ 时，或许不能计算出 $\mathbf{g}_1$ 或 $\mathbf{A}_1$。

当目标函数不是全局凹的时候，一种好的实践做法是，使用一系列初始值，增大得到全局最大值的机会。

10.2.8 *数值导数与解析导数*

由定义知，任何梯度法都使用目标函数的导数。或者使用数值导数，或者使用解析导数。

数值导数(numberical derivatives)是利用:

$$\frac{\Delta Q_N(\hat{\boldsymbol{\theta}}_s)}{\Delta \theta_j}=\frac{Q_N(\hat{\boldsymbol{\theta}}_s+h\mathbf{e}_j)-Q_N(\hat{\boldsymbol{\theta}}_s-h\mathbf{e}_j)}{2h},\quad j=1,\cdots,q \tag{10.4}$$

计算,其中,h 很小,而 $\mathbf{e}_j=(0\cdots 0\ 1\ 0\cdots 0)'$表示第 j 行为 1 而其余行为 0 的向量。

从理论上讲,h 应该是非常小的,因为当 $h\to 0$ 时,正式讲,$\Delta Q_N(\boldsymbol{\theta})/\Delta\theta_j$ 等于 $\partial Q_N(\boldsymbol{\theta})/\partial\theta_j$ 的极限。在实际应用时,太小的 h 值会导致不准确,其原因在于舍入误差。正因为这个缘由,利用数值导数的计算总是应当采取双倍精度或四倍精度,而不是单精度的。尽管程序使用默认值,比如 $h=10^{-6}$,但对于特殊问题来说,其他值将会更好。例如,若 NLS 回归中的因变量 y 以千美元来计量,而不是以美元来计量(回归元没有重新标度),很小的 h 值就适宜,从而,$\boldsymbol{\theta}$ 将是千分之一的大小。

利用数值导数的缺点是,对于 q 个参数的每一个、N 个观测值中的每一个以及 S 次迭代中的每一次,这些导数必须计算多次。这要求对目标函数计算 $2qNS$ 次,其中的每一次计算值在计算形式上或许是繁琐而艰难的。

一种可供选择的方法是,使用解析导数(**analytical derivatives**)。和数值导数相比,这些方法将更准确,并计算起来会更快捷,尤其是解析导数比其目标函数本身的计算更简单。此外,只需要进行 qNS 次的函数计算。

对于额外需要计算二次导数来建立 $\mathbf{A}_s$ 的方法来说,甚至在提供解析导数方面,存在着较大优势。即使一阶解析导数只是给定的,二阶导数也会更迅速而准确地成为一阶解析导数的一阶数值导数。统计软件经常为用户提供一阶与二阶解析导数的选项。

一些数值导数,除了提供目标函数之外,具有不需要编程的优点。这可以节省编程时间并剔除用户可能的错误来源,尽管某些软件有能力计算解析导数。

不过,假如计算时间是一个因素,或者关切计算的准确性,则特别提供解析导数是值得的。于是,一种好的实践做法是,检验解析导数可通过利用数值导数获得参数估计值来正确地对解析导数编程,并且利用解析导数获得的初始值估计值。

10.2.9 *非梯度方法*

为了确保梯度存在,梯度法假定目标函数是充分光滑的。有一些例子,比如著名的最小绝对偏差(LAD)、分位数回归以及最大得分估计,可使用一种非梯度的且可供选择的迭代法。

例如,对于 LAD 来说,目标函数 $Q_N(\boldsymbol{\theta}_s)=N^{-1}\sum_i|y_i-\mathbf{x}_i\boldsymbol{\beta}|$ 不存在导数,从而运用线性规划方法(**linear programming methods**)代替梯度法。这类例子,在我们特别关注于梯度法的微观经济计量学中几乎十分少见。

对于很难求最大值的目标函数,尤其是由于出现多重局部最优值,要使用一些非梯度法,诸如模拟退火法(10.3.8 节将阐述)与遗传算法(**genetic algorithms**)[参见多尔西和迈耶(Dorsey and Mayer, 1995)]。

10.3 特定方法

最大化全局凹的目标函数的一个重要方法是牛顿—拉夫森迭代法。当牛顿—拉夫森方法失效时，其他一些方法，诸如最速下降法以及 DFP，通常都是人们想学习并利用的。对于 NLS 估计量来说，另一种常用方法是高斯—牛顿方法。该方法虽不像牛顿—拉夫森方法那样通用，但它只有对最小二乘问题才可应用，而且它可作为对牛顿—拉夫森方法的稍微修改。人们将这些各种各样的方法设计成用于获得给定参数的具有某些初始值的局部最优值。

本节还将阐述期望值方法，它尤其对缺失数据问题有用，而模拟退火法则是非梯度法的一个例子，并且该方法最有可能产生全局最大值而不是局部最大值。

10.3.1 *牛顿—拉夫森方法*

牛顿—拉夫森方法[**Newton - Raphson (NR) method**]是一种十分流行的梯度法。倘若目标函数关于 $\boldsymbol{\theta}$ 是全局凹的，则此方法特别有效。在这一方法中：

$$\hat{\boldsymbol{\theta}}_{s+1}=\hat{\boldsymbol{\theta}}_s-\mathbf{H}_s^{-1}\mathbf{g}_s \tag{10.5}$$

其中，$\mathbf{g}_s$ 已由式(10.2)定义，而：

$$\mathbf{H}_s=\left.\frac{\partial^2 Q_N(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\hat{\theta}_s} \tag{10.6}$$

表示 $q\times q$ 阶海赛矩阵在 $\hat{\boldsymbol{\theta}}_s$ 处的计算值。这些公式既可用于求 $Q_N(\boldsymbol{\theta})$ 的最大值，又可用于求最小值，因为用负号乘以 $Q_N(\boldsymbol{\theta})$，人们既能改变 $\mathbf{H}_s^{-1}$ 的符号，又能改变 $\mathbf{g}_s$ 的符号。

为引出牛顿—拉夫森(NR)方法，以关于 $\boldsymbol{\theta}$ 的第 s 次估计值 $\hat{\boldsymbol{\theta}}_s$ 开始。然后，借助于在 $\hat{\boldsymbol{\theta}}_s$ 处的二阶泰勒级数展开：

$$Q_N(\boldsymbol{\theta})=Q_N(\hat{\boldsymbol{\theta}}_s)+\left.\frac{\partial Q_N(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right|_{\hat{\theta}_s}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_s)+\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_s)'\left.\frac{\partial^2 Q_N(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\hat{\theta}_s}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_s)+R$$

一旦忽略余项 R，同时利用更简洁记号，我们通过

$$Q_N^*(\boldsymbol{\theta})=Q_N^*(\hat{\boldsymbol{\theta}}_s)+\mathbf{g}_s'(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_s)+\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_s)'\mathbf{H}_s(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_s)$$

逼近 $Q_N(\boldsymbol{\theta})$，其中，$\mathbf{g}_s$ 与 $\mathbf{H}_s$ 都已在式(10.2)与式(10.6)中定义。为近似求出 $Q_N^*(\boldsymbol{\theta})$ 关于 $\boldsymbol{\theta}$ 的最大值，我们令其导数为 0。于是，$\mathbf{g}_s+\mathbf{H}_s(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_s)=\mathbf{0}$，并求解 $\boldsymbol{\theta}$，得到 $\hat{\boldsymbol{\theta}}_{s+1}=\hat{\boldsymbol{\theta}}_s-\mathbf{H}_S^{-1}$，这就是式(10.5)。因此，NR 最大化二阶泰勒级数对在 $\hat{\boldsymbol{\theta}}_s$ 处估计的 $Q_N(\boldsymbol{\theta})$ 的逼近。

为了理解 NR 迭代是否一定使 $Q_N(\boldsymbol{\theta})$ 增大，把第($s+1$)次估计值代入泰勒级数近似式中，得到：

$$Q_N(\hat{\boldsymbol{\theta}}_{s+1})=Q_N(\hat{\boldsymbol{\theta}}_s)-\frac{1}{2}(\hat{\boldsymbol{\theta}}_{s+1}-\hat{\boldsymbol{\theta}}_s)'\mathbf{H}_s(\hat{\boldsymbol{\theta}}_{s+1}-\hat{\boldsymbol{\theta}}_s)+R$$

一旦忽略余项，可以发现，当 $\mathbf{H}_s$ 是负定的（或正定的），$Q_N(\hat{\boldsymbol{\theta}}_{s+1})$ 就将增大（或减少）。在局部最大值处，海赛矩阵是半负定的，但离开最大值时，甚至对于定义良好的问题来说，或许不是这种情况。如果 NR 方法误入这种领域，那么它不一定朝最大值运动。进一步地，当海赛矩阵是奇异的时，不能计算式（10.5）中的 $\mathbf{H}_S^{-1}$。很明显，如果目标函数是全局凹的（或者凸的），进而 $\mathbf{H}_s$ 总是负定的（或正定的），那么 NR 方法对最大化问题（或者最小化问题）最为有效。在这些情况下，收敛时常在 10 次迭代之内出现。

如果初始值 $\hat{\boldsymbol{\theta}}_1$ 是根号 N 一致的，也就是说，如果 $\sqrt{N}(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_0)$ 服从正常极限分布，NR 方法就具有额外引人注目的特征。于是，可以证明，第二次估计量 $\hat{\boldsymbol{\theta}}_2$ 具有与通过迭代而获得收敛估计量一样的渐近分布。因此，进一步迭代并不会在理论上得到好处。一个例子是可行的 GLS，其中，最初 OLS 导致一致回归参数估计，而这些值同样用于获得一致方差参数估计，进而用于获得有效 GLS。第二个例子是，在对复杂的似然函数求最大值之前，运用很容易获得的一致估计值作为初始值。尽管不要求做进一步迭代，但在实际应用中，研究者还是喜欢通过迭代来达到收敛，除非这样做在计算上太耗费时间。迭代收敛的一个优点是，不同的研究者应获得相同的参数估计值，而各种不同的根号 N 一致估计会导致第二次参数估计值各不相同，尽管它们都是渐近等价的。

10.3.2 *得分方法*

一种对 NR 方法进行普遍修改的方法是得分方法（**method of scoring**）。在这个方法中，海赛矩阵要用以下期望值来代替：

$$\mathbf{H}_{MS,s}=\mathrm{E}\left[\frac{\partial^2 Q_N(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]\bigg|_{\hat{\theta}_s} \tag{10.7}$$

此种代换用于 MLE［也就是说，当 $Q_N(\boldsymbol{\theta})=N^{-1}\mathcal{L}_N(\boldsymbol{\theta})$］时特别有利，因为由信息矩阵等式(参见 5.6.3 节)可知，期望值应是负定的，$\mathbf{H}_{MS,s}=-\mathrm{E}[\mathcal{L}_N/\partial\boldsymbol{\theta}\partial\mathcal{L}_N/\partial\boldsymbol{\theta}']$[1]，由于它是一个协方差矩阵，所以 E［·］是正定的。对 m 估计量来说，只有获得式（10.7）中的期望才是可行的，而且甚至在此情况下使用解析方法或许都很难。

对于广义线性模型的 MLE 来说，诸如泊松、probit 以及 logit，可以证明，利用迭代重新加权最小二乘法的得分方法是可行的［参见麦卡拉和内尔德（McCullagh and Nelder, 1989)］。这有利于及早采用只可使用 OLS 程序的模型。

得分方法还应用于 m 估计量，而不是 MLE，尽管 $\mathbf{H}_{MS,s}$ 可能不是负的。

10.3.3 *BHHH 方法*

伯思特、霍尔、霍尔以及豪斯曼（Berndt, Hall, Hall and Hausman, 1974）的 BHHH 方法（**BHHH method**）利用加权矩阵 $\mathbf{A}_s=-\mathbf{H}_{BHHH,s}^{-1}$ 来使用式（10.1），其中：

$$\mathbf{H}_{BHHH,s}=-\sum_{i=1}^{N}\frac{\partial q_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial q_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\bigg|_{\hat{\theta}_s} \tag{10.8}$$

---

〔1〕 原著中这个式子没有负号，但应添加负号。——译者注

而 $Q_N(\boldsymbol{\theta})=\sum_i q_i(\boldsymbol{\theta})$。和 NR 相比，这具有仅需计算一阶导数值的优点，所以大大简化了计算量。

为使该方法正确，以 MLE 的得分方法开始，在此情况下，$Q_N(\boldsymbol{\theta})=\sum_i \ln f_i(\boldsymbol{\theta})$，其中，$f_i(\boldsymbol{\theta})$表示对数密度。信息矩阵等式写成：

$$\mathrm{E}\left[\frac{\partial^2 \mathcal{L}_N(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]=-\mathrm{E}\left[\sum_{i=1}^{N}\frac{\partial \ln f_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\sum_{j=1}^{N}\frac{\partial \ln f_j(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right]$$

而且对于不同 i 的独立性蕴含：

$$\mathrm{E}\left[\frac{\partial^2 \mathcal{L}_N(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]=-\sum_{i=1}^{N}\mathrm{E}\left[\frac{\partial \ln f_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial \ln f_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right]$$

一旦省掉期望，就得到式(10.8)。

BHHH 方法还能用于一些估计量而不是 MLE 上，在此情况下，它被看成是对式(10.1)中矩阵 $\mathbf{A}_s$ 的另一种直接选取，而不是对海赛矩阵 $\mathbf{H}_s$ 的估计。

BHHH 方法可用于许多横截面 m 估计量上，因为它非常有效且只需要一阶导数。

10.3.4 *最速下降法*

最速下降法(**method of steepest ascent**)是设 $\mathbf{A}_s=\mathbf{I}_q$，即对加权矩阵进行最简单选取。于是，线搜索是通过常值 λ_s 来标度 $\mathbf{I}_q$ 来实施的[参见式(10.3)]。

线搜索是以手工操作形式下降的。在实际应用中，一种普遍做法是运用线搜索的最优 λ，可以证明，$\lambda_s=-\mathbf{g}_s'\mathbf{g}_s/\mathbf{g}_s'\mathbf{H}_s\mathbf{g}_s$，其中，$\mathbf{H}_s$ 表示海赛矩阵，这个最优 λ_s 需要计算海赛矩阵，在此情况下，人们反而要使用 NR。作为最速下降而不是 NR 的优点是，$\mathbf{H}_s$可能是奇异的，尽管还需要 $\mathbf{H}_s$是负定的来确保 $\lambda_s<0$，因此，$\lambda_s\mathbf{I}_q$ 是负定的。

10.3.5 *DFP 与 BFGS 方法*

DFP 算法(**DFP algorithm**)归功于达维登(Davidon)、弗莱彻(Fletcher)和鲍威尔(Pouell)，该方法是含有正定加权矩阵 $\mathbf{A}_s$ 的一种梯度法，并只需要计算一阶导数，而不像 *NR* 需要计算海赛矩阵。这里，对此方法只阐述而没有推导。

加权矩阵 $\mathbf{A}_s$ 可通过递归：

$$\mathbf{A}_s=\mathbf{A}_{s-1}+\frac{\boldsymbol{\delta}_{s-1}\boldsymbol{\delta}_{s-1}'}{\boldsymbol{\delta}_{s-1}'\boldsymbol{\gamma}_{s-1}}+\frac{\mathbf{A}_{s-1}\boldsymbol{\gamma}_{s-1}\boldsymbol{\gamma}_{s-1}'\mathbf{A}_{s-1}}{\boldsymbol{\gamma}_{s-1}'\mathbf{A}_{s-1}\boldsymbol{\gamma}_{s-1}} \tag{10.9}$$

其中，$\boldsymbol{\delta}_{s-1}=\mathbf{A}_{s-1}\mathbf{g}_{s-1}$，$\boldsymbol{\gamma}_{s-1}=\mathbf{g}_s-\mathbf{g}_{s-1}$，通过对式(10.9)的右边进行检查，倘若初始 $\mathbf{A}_0$ 是正定的(比如，$\mathbf{A}_0=\mathbf{I}_q$)，则 $\mathbf{A}_s$将是正定的。

在许多统计应用中，此程序很快就收敛。最后，$\mathbf{A}_s$ 趋于理论上偏爱的 $-\mathbf{H}_s^{-1}$。原则上，该方法还能提供用于计算标准误差中海赛矩阵的近似估计，而不需要二次导数或者矩阵逆。可是，在实际应用时，这种估计是一个不好的估计。

对 DFP 算法的精炼是 BFGS 算法(**BFGS algorithm**)，此方法中，博伊登(Boyden)、弗莱彻(Fletchor)、戈德法布(Goldfarb)和香农(Shannon)使用：

$$\mathbf{A}_s=\mathbf{A}_{s-1}+\frac{\boldsymbol{\delta}_{s-1}\boldsymbol{\delta}_{s-1}'}{\boldsymbol{\delta}_{s-1}'\boldsymbol{\gamma}_{s-1}}+\frac{\mathbf{A}_{s-1}\boldsymbol{\gamma}_{s-1}\boldsymbol{\gamma}_{s-1}'\mathbf{A}_{s-1}}{\boldsymbol{\gamma}_{s-1}'\mathbf{A}_{s-1}\boldsymbol{\gamma}_{s-1}}-(\boldsymbol{\gamma}_{s-1}'\mathbf{A}_{s-1}\boldsymbol{\gamma}_{s-1})\boldsymbol{\eta}_{s-1}\boldsymbol{\eta}_{s-1}' \tag{10.10}$$

其中，$\boldsymbol{\eta}_{s-1}=(\boldsymbol{\delta}_{s-1}/\boldsymbol{\delta}_{s-1}'\boldsymbol{\gamma}_{s-1})-(\mathbf{A}_{s-1}\boldsymbol{\gamma}_{s-1}/\boldsymbol{\gamma}_{s-1}'\mathbf{A}_{s-1}\boldsymbol{\gamma}_{s-1})$。

10.3.6 *高斯—牛顿方法*

高斯—牛顿方法[**Gauss-Newton(GN) method**]是关于 MLS 估计量的迭代法，而 NLS 估计量能通过迭代 OLS 来执行。

明确地讲，对于含有条件均值函数 $g(\mathbf{x}_i,\boldsymbol{\beta})$的 NLS 来说，GN 方法是设参数变化向量$(\hat{\boldsymbol{\beta}}_{s+1}-\hat{\boldsymbol{\beta}}_s)$等于源于人工回归

$$y_i-g(\mathbf{x}_i,\hat{\boldsymbol{\beta}}_s)=\left.\frac{\partial g_i}{\partial \boldsymbol{\beta}'}\right|_{\hat{\boldsymbol{\beta}}_s}\boldsymbol{\beta}+v_i \tag{10.11}$$

的 OLS 系数估计值。等价地，$\hat{\boldsymbol{\beta}}_{s+1}$等于源自人工回归

$$y_i-g(\mathbf{x}_i,\hat{\boldsymbol{\beta}}_s)-\left.\frac{\partial g_i}{\partial \boldsymbol{\beta}'}\right|_{\hat{\boldsymbol{\beta}}_s}\hat{\boldsymbol{\beta}}_s=\left.\frac{\partial g_i}{\partial \boldsymbol{\beta}'}\right|_{\hat{\boldsymbol{\beta}}_s}\boldsymbol{\beta}+v_i \tag{10.12}$$

的 OLS 系数估计值。

为了推导这种方法，令 $\hat{\boldsymbol{\beta}}_s$ 是一个初始值，通过一阶泰勒级数展开式

$$g(\mathbf{x}_i,\boldsymbol{\beta})=g(\mathbf{x}_i,\hat{\boldsymbol{\beta}}_s)+\left.\frac{\partial g_i}{\partial \boldsymbol{\beta}'}\right|_{\hat{\boldsymbol{\beta}}_s}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_s)$$

来逼近 $g(\mathbf{x}_i,\boldsymbol{\beta})$，并把它代入最小二乘法目标函数 $Q_N(\boldsymbol{\beta})$，得到近似：

$$Q_N^*(\boldsymbol{\beta})=\sum_{i=1}^N\left(y_i-g(\mathbf{x}_i,\hat{\boldsymbol{\beta}}_s)-\left.\frac{\partial g_i}{\partial \boldsymbol{\beta}}\right|_{\hat{\boldsymbol{\beta}}_s}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_s)\right)^2$$

这是 $y_i-g(\mathbf{x}_i,\hat{\boldsymbol{\beta}}_s)$对含有参数向量$(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_s)$的$\partial g_i/\partial\boldsymbol{\beta}'|_{\hat{\boldsymbol{\beta}}_s}$进行 OLS 回归的残差平方和，由此得出式(10.11)。更正式地讲：

$$\hat{\boldsymbol{\beta}}_{s+1}=\hat{\boldsymbol{\beta}}_s+\left[\sum_i\left.\frac{\partial g_i}{\partial \boldsymbol{\beta}}\right|_{\hat{\boldsymbol{\beta}}_s}\left.\frac{\partial g_i}{\partial \boldsymbol{\beta}'}\right|_{\hat{\boldsymbol{\beta}}_s}\right]^{-1}\sum_i\left.\frac{\partial g_i}{\partial \boldsymbol{\beta}}\right|_{\hat{\boldsymbol{\beta}}_s}(y_i-g(\mathbf{x}_i,\hat{\boldsymbol{\beta}}_s)) \tag{10.13}$$

这就是含有向量 $\mathbf{g}_s=\sum_i\partial g_i/\partial\boldsymbol{\beta}|_{\hat{\boldsymbol{\beta}}_s}(y_i-g(\mathbf{x}_i,\hat{\boldsymbol{\beta}}_s))$并通过矩阵 $\mathbf{A}_s=[\sum_i\partial g_i/\partial\boldsymbol{\beta}\times\partial g_i/\partial\boldsymbol{\beta}'|_{\hat{\boldsymbol{\beta}}_s}]^{-1}$加权的梯度法(10.1)。

由 5.8 节知，迭代法(10.13)等于 NLS 估计的牛顿—拉夫森算法的得分变形方法，其右边的第二项和是梯度向量，而第一项和是负的海赛矩阵期望值(还可参见 10.3.9 节)。因此，高斯—牛顿算法是牛顿—拉夫森的一种特殊情况，此处更强调 NR，因为它与 GN 相比能够应用于更加广泛的问题。

10.3.7 *期望最大化*

可以认为，本书考察的众多数据与模型公式都涉及不完整数据及缺失数据。例如，关注的结果变量(比如，某个州在某个时期的开支或时期长度)可能是右删失。也就是说，在一些情况下，我们可观测到真实开支或时期长度，而在另一些情

况下，我们可能只知道结果大于某一个特定值，比如说 c^*。第二个例子涉及多元回归，其数据矩阵看起来如下：

$$\begin{bmatrix} \mathbf{y}_1 & \mathbf{X}_1 \\ ? & \mathbf{X}_2 \end{bmatrix}$$

其中，? 代表缺失数据。此外，我们面临下述情况：想要估计线性回归模型 $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{u}$，其中 $\mathbf{y}'=[\mathbf{y}_1\quad ?]$，$\mathbf{X}'=[\mathbf{X}_1\quad \mathbf{X}_2]$，但变量 $\mathbf{y}$ 的某个子集出现缺失。第三个例子涉及估计 C 个成分混合分布的参数($\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\cdots,\boldsymbol{\theta}_C,\pi_1,\cdots,\pi_C$)，该例子还被称为潜类模型，$h(y|\mathbf{X})=\sum_{j=1}^{C}\pi_j f_j(\mathbf{y}_j|\mathbf{X}_j,\boldsymbol{\theta}_j)$，其中，$f_j(\mathbf{y}_j|\mathbf{X}_j,\boldsymbol{\theta}_j)$表示定义良好的 pdf。这里，$\pi_j(j=1,\cdots,C)$表示未知的抽样部分，其对应于源自抽样的观测值 C 潜类密度。如果抽样部分是已知常值，那么估计起来就比较简单，在这个意义上，把此问题看成是缺失数据问题会很方便。

期望最大化(**expectation maximization，EM**)框架提供用于能被解释成涉及缺失数据问题的发展算法的框架。关于对这种类型的估计问题进行特殊求解的文献由来已久，但登普斯特、莱尔德和鲁宾(Dempster，Laird，and Rubin，1977)却提供了最可靠的处理。

设 $\mathbf{y}$ 表示关注因变量变量，由基本的潜变量向量 $\mathbf{y}^*$ 来决定。设 $f^*(\mathbf{y}^*|\mathbf{X},\boldsymbol{\theta})$表示以回归元 $\mathbf{X}$ 为条件的潜变量的联合密度，并设 $f(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})$表示已观测到变量的联合密度。设从 $\mathbf{y}$ 的样本空间到 $\mathbf{y}^*$ 的空间存在多对一映射，也就是说，潜变量 $\mathbf{y}^*$ 的值唯一地决定 $\mathbf{y}$，但 $\mathbf{y}$ 的值并不唯一地决定 $\mathbf{y}^*$。由此可得，$f(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})=f^*(\mathbf{y}^*|\mathbf{X},\boldsymbol{\theta})/f(\mathbf{y}^*|\mathbf{y},\mathbf{X},\boldsymbol{\theta})$，因为由贝叶斯规则条件密度 $f(\mathbf{y}^*|\mathbf{y})=f(\mathbf{y},\mathbf{y}^*)/f(\mathbf{y})=f^*(\mathbf{y}^*)/f(\mathbf{y})$，其中，最后等式使用了 $f(\mathbf{y}^*,\mathbf{y})=f^*(\mathbf{y}^*)$ 作为 $\mathbf{y}^*$ 唯一地决定 $\mathbf{y}$。重新排列得到，$f(\mathbf{y})=f^*(\mathbf{y}^*)/f^*(\mathbf{y}^*|\mathbf{y})$。

MLE 对

$$Q_N(\boldsymbol{\theta})=\frac{1}{N}\mathcal{L}_N(\boldsymbol{\theta})=\frac{1}{N}\ln f^*(\mathbf{y}^*|\mathbf{X},\boldsymbol{\theta})-\frac{1}{N}\ln f(\mathbf{y}^*|\mathbf{y},\mathbf{X},\boldsymbol{\theta}) \qquad (10.14)$$

求最大值。因为 $\mathbf{y}^*$ 是不可观测的，所以对数似然中的第一项可被忽略掉。第二项用它自己的期望值来代替，这样做将不涉及 $\mathbf{y}^*$，在第 s 次，该期望要在 $\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_s$ 处计算。

EM 算法(**EM algorithm**)的期望(E)要计算：

$$Q_N(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}_s)=-\mathrm{E}\left[\frac{1}{N}\ln f(\mathbf{y}^*|\mathbf{y},\mathbf{X},\boldsymbol{\theta})\,|\,\mathbf{y},\mathbf{X},\hat{\boldsymbol{\theta}}_s\right] \qquad (10.15)$$

其中，期望是关于密度 $f(\mathbf{y}^*|\mathbf{y},\mathbf{X},\hat{\boldsymbol{\theta}}_s)$的。EM 算法的最大化[**maximization** (M)]部分是求 $Q_N(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}_s)$的最大值，从而获得 $\hat{\boldsymbol{\theta}}_{s+1}$。

完全 EM 算法是一种迭代。已知潜变量的期望值，对似然函数求最大值；已知 $\boldsymbol{\theta}$ 的当前值，重新计算期望值。该迭代过程连续不断地进行，一直到收敛为止。EM 算法具有总使 $Q_N(\boldsymbol{\theta})$增大或者恒定的优点，参见雨宫(Amemiya，1985，第 376 页)。EM 算法在 18.5.3 节用于潜类模型，而在 27.5 节则用于缺失数据。

尽管 EM 算法只能用于最优化问题的一个子集，但在 EM 算法有效应用方面存在着相当广泛的文献。在许多情况下，EM 算法容易编程，并且对它的使用进一步地受到有限计算能力与内存考虑的激发，有限计算能力与内存现在已不再是最高的。虽然 EM 算法具有这些吸引力。但对于删失数据模型与潜类型来说，直接利用牛顿—拉夫森类型迭代程序的估计时，常发现计算会更快一些且更有效。

10.3.8 *模拟退火*

模拟退火（**Simulated Annealing**, SA）是另一种非梯度迭代法，戈夫、费里尔和罗杰斯（Goffe, Ferrier, and Rogers, 1994）对该方法给出一个综述。它允许目标函数向减少方向而不是向增大方向运动，不同于梯度法，因而它没有锁定向着其特殊局部最大值的稳定运动。

已知第 s 次迭代值 $\hat{\boldsymbol{\theta}}_s$，我们扰动 $\hat{\boldsymbol{\theta}}_s$ 的第 j 分量，获得一个新的试验值：

$$\boldsymbol{\theta}_s^* = \hat{\boldsymbol{\theta}}_s + [0 \cdots 0 \ (\lambda_j r_j) \ 0 \cdots 0]' \tag{10.16}$$

其中，λ_j 表示预先设定的步长，而 r_j 表示从（-1, 1）均匀分布中所抽取的。当使用新的试验值时，也就是说，此方法设 $\hat{\boldsymbol{\theta}}_{s+1} = \boldsymbol{\theta}_s^*$，这样做要么使目标函数增大，要么没有使目标函数值增大，却通过了梅特罗波利斯准则（**Metropolis criterion**）：

$$\exp((Q_N(\boldsymbol{\theta}_s^*) - Q_N(\hat{\boldsymbol{\theta}}_s))/T_s) > u \tag{10.17}$$

其中，u 表示从（0, 1）均匀分布中所抽取的，而 T_s 表示被称为温度（**temperature**）的标度参数。因而，对于使 $Q_N(\boldsymbol{\theta}_s^*)$ 与 $Q_N(\hat{\boldsymbol{\theta}}_s)$ 之差减少的概率且促使温度增高来说，不仅上升运动可被接收，而且下降运动同样被接收。术语模拟退火与温度均起源于，与通过缓慢冷却浇铸金属的最小化热能理论的类比。

使用者需要设定步长参数 λ_j。戈夫（Goffe, 1994）对 λ_j 进行周期性调整，以使所有一系列迭代的运动中有 50% 是可接收的。同理，需要对温度加以选择，从而减少迭代过程。于是，此算法在稳定地锁定一个特殊区域之前，最初要在广泛的参数范围内进行搜寻。

快速模拟退火（**fast simulated annealing**, FSA）是一种较快的方法，它是由舒和哈特利（Szu and Hartley, 1987）提出的。它是用由温度标度的柯西随机变量 v_j 代替（-1, 1）均匀分布随机数 r_j，并允许固定步长 v_j。该方法还使用 T_s 等于初始温度被 FSA 迭代次数除，对迭代温度进行较简单的调整，其中一次迭代就是对 $\boldsymbol{\theta}$ 的第 q 个分量的一个完整循环。

卡梅伦和约翰森（Cameron and Johansson, 1997）沿着霍罗维茨（Horowitz, 1992）的方法对模拟退火加以讨论并使用。这以 FSA 开始，当在一系列迭代或众多次（250 次）迭代之后，$Q_N(\cdot)$ 上产生相对很小的变化，出于节省计算量的缘故，转换到梯度法（BFGS）上。在模拟研究时，他们发现，与仅含有一个初始值的 NR 相比，含有一系列不同初始值的 NR 得到了相当大的改进，但还有更好的，那就是含有一系列不同初始值的 FSA。

10.3.9 *例子：指数回归*

考察含有指数条件均值

$$E[y_i|\mathbf{x}_i]=\exp(\mathbf{x}_i'\boldsymbol{\beta}) \tag{10.18}$$

的非线性回归模型，其中，$\mathbf{x}_i$ 与 $\boldsymbol{\beta}$ 均表示 $K\times 1$ 维向量。NLS 估计量 $\hat{\boldsymbol{\beta}}$ 是

$$Q_N(\boldsymbol{\beta})=\sum_i(y_i-\exp(\mathbf{x}_i'\boldsymbol{\beta}))^2 \tag{10.19}$$

的极小值解，其中，为了记号简单，忽略了通过 $2/N$ 进行的标度。其一阶条件关于 $\boldsymbol{\beta}$ 是非线性的，同时 $\boldsymbol{\beta}$ 没有显式解。相反，需要使用梯度法。

就这个例子而言，梯度与海赛矩阵分别是：

$$\mathbf{g}=-2\sum_i(y_i-e^{\mathbf{x}_i'\boldsymbol{\beta}})e^{\mathbf{x}_i'\boldsymbol{\beta}}\mathbf{x}_i \tag{10.20}$$

与

$$\mathbf{H}=2\sum_i\{e^{\mathbf{x}_i'\boldsymbol{\beta}}e^{\mathbf{x}_i'\boldsymbol{\beta}}\mathbf{x}_i\mathbf{x}_i'-2(y_i-e^{\mathbf{x}_i'\boldsymbol{\beta}})e^{\mathbf{x}_i'\boldsymbol{\beta}}\mathbf{x}_i\mathbf{x}_i'\} \tag{10.21}$$

NR 迭代法(10.5)使用了式(10.20)与式(10.21)在 $\hat{\boldsymbol{\beta}}_s$ 处的计算值 $\mathbf{g}_s$ 与 $\mathbf{H}_s$。

注意式(10.18)，NR 的一种较简单的得分变形蕴含：

$$E[\mathbf{H}]=2\sum_i e^{\mathbf{x}_i'\boldsymbol{\beta}}e^{\mathbf{x}_i'\boldsymbol{\beta}}\mathbf{x}_i\mathbf{x}_i' \tag{10.22}$$

一旦利用 $E[\mathbf{H}_s]$ 代替 $\mathbf{H}_s$，就得到：

$$\hat{\boldsymbol{\beta}}_{s+1}-\hat{\boldsymbol{\beta}}_s=\left[\sum_i e^{\mathbf{x}_i'\hat{\boldsymbol{\beta}}_s}e^{\mathbf{x}_i'\hat{\boldsymbol{\beta}}_s}\mathbf{x}_i\mathbf{x}_i'\right]^{-1}\sum_i e^{\mathbf{x}_i'\hat{\boldsymbol{\beta}}_s}\mathbf{x}_i(y_i-e^{\mathbf{x}_i'\hat{\boldsymbol{\beta}}_s})$$

由此可得，从 $(y_i-e^{\mathbf{x}_i'\hat{\boldsymbol{\beta}}_s})$ 对 $e^{\mathbf{x}_i'\hat{\boldsymbol{\beta}}_s}\mathbf{x}_i$ 的 OLS 回归中计算 $\hat{\boldsymbol{\beta}}_{s+1}-\hat{\boldsymbol{\beta}}_s$。对于指数条件均值(10.18)来说，由于 $\partial g(\mathbf{x}_i,\boldsymbol{\beta})/\partial\boldsymbol{\beta}=\exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}}_s)\mathbf{x}_i$，这也是高斯—牛顿回归(10.11)。对 $\exp(\mathbf{x}_i'\boldsymbol{\beta})=\exp(\boldsymbol{\beta})$ 专门研究，可得到 10.2.4 节曾阐述的迭代程序。

10.4 应用研究

一些实际应用问题已经在 10.2 节阐述过，包括著名的收敛准则，诸如步长的调整，还有对数值导数而不是解析导数的利用。本节将对一些统计软件包给予简要概述，并对沿着非线性估计量计算时易犯的普通错误展开讨论。

10.4.1 *统计软件*

所有标准的微观经济计量学软件包，比如 Limdep、Stata、PCTSP 以及 SAS 都具有嵌入式程序，这些都能估计基本的非线性模型，诸如 logit 与 probit。一旦不需要迭代或甚至所用模型的知识，就可直接运用这些软件包。例如，关于 logit 回归的命令可以是“logit y z”而不是关于 OLS 的命令“ols y z”。非线性最小二乘法需要某种编程，以便包含人们希望设定 $g(\mathbf{x},\boldsymbol{\beta})$ 的特殊函数形式。由于程序应该利用特殊模型的结构，所以估计应很快捷且准确。例如，如果目标数是全局凹的，就可利用得分方法。

倘若统计软件包没有包括特殊模型，则要求人们编写自己的特有程序。这种

情况甚至对于标准模型的稍微变形，比如对一些参数施加约束或利用不具有单指标形式的参数化，都会出现。人们编程可利用特别喜欢的统计软件包或其他一些更专门的编程语言来编写。一些可能情况包括：(1) 嵌入在统计软件包之中的最优化程序需要对目标函数及其可能的导数加以设定；(2) 统计软件包中的矩阵命令要计算 $\mathbf{A}_s$ 与 $\mathbf{g}_s$，以及迭代；(3) 矩阵编程语言，诸如 Gauss、Matlab、OX、SAS/IML 或者 S-Plus，都可能附有最优化程序；(4) 编程语言，诸如 Fortran 或 C++；(5) 最优化软件包，诸如 GAMS、GQOPT 或 NAGLIB 中的那些最优化部分。

第一种方法与第二种方法颇具吸引力，因为它们不需要用户学习新的程序。对 m 估计来说，第一种方法尤其简单，原因在于该方法只要求对第 i 个观测值的子函数 $q_i(\boldsymbol{\theta})$ 加以设定，而不是对 $Q_N(\boldsymbol{\theta})$ 进行设定。可是，在实际应用中，标准软件包中关于用户可定义函数的最优化程序与使用更专门化程序相比，最可能遇到数值问题。此外，对一些软件包而言，第二种方法就要求学习深奥难懂的矩阵编程形式。

对于非线性问题来说，第三种方法最好，尽管这要求用户从头学习矩阵编程语言。而且，实际上人们开始处理任何遇到的经济计量问题，一般地讲，含有矩阵编程语言的最优化程序是适宜的。再者，一些作者还运用特定论文中曾用过的程序。

一般地讲，第四种方法与第五种方法要求用户具有比第三种方法更为复杂的编程水平。第四种方法能产生更快速的计算，而第五种方法则能解决在数值计算上最具挑战意义的最优化问题。

另外，其他一些实际问题包括软件成本；同事使用何种软件；一个软件是否具有清楚的误差信息以及有益的排除程序错误的特性，比如逐行追踪程序执行的跟踪程序。运用类似于由其他同事所用的软件价值不能被低估。

10.4.2 *计算困难*

在实际应用中，计算困难在于不可能获得参数估计。例如，误差信息可以表明无法计算估计量，因为海赛矩阵是奇异的。出现此情况存在许多可能原因，如同表 10.2 所概述的。这些原因还给出了对参数估计的另一种普遍情况的解释，其中参数估计显然处于误差之中。

表 10.2 计算困难：实际核查项目

问题	核查内容
不正确读取数据	打印出全部描述统计量
不准确的计算	使用含有各种步长 h 的不同解析导数或数值导数
多重共线性	核查 $\mathbf{X}'\mathbf{X}$ 的条件数。尝试回归元子集
迭代出现奇异矩阵	尝试不需要矩阵逆的方法，比如 DFP
不好的初始值	尝试一系列不同的初始值
模型不可识别	核查起来困难。明显的核查是虚拟变量陷阱
奇怪的参数值	是包括还是排除常值？迭代确实收敛吗？
不同标准误差	用哪一种方法计算方差矩阵？

第一,数据没有以正确方式读入。这是一种值得注意的普遍出错的类型。对于大的数据集来说,用打印机打印出所有数据是不切实际的。不过,人们至少应该总是获得描述统计量,并核对一些异常现象,诸如变量的不正确范围,是否有异乎寻常大的均值或异乎寻常小的均值,以及异乎寻常大的或小的标准差(包括零值,这表示没有变异)。更详细内容参见 3.5.4 节。

第二,可能存在一些计算误差。为使这些误差减到最小范围,所有计算都应以 2 倍精度甚至 4 倍精度而不是单精度进行运算。对数据重新标度是有益的。因而,回归元具有类似的均值与方差。例如,或许更好的方式是使用千美元测量年收入,而不是以美元为单位计量的。若使用数值导数,就必须对式(10.4)中变化值 h 加以改动,需要关注的是,如何计算函数值。例如,最好利用对数伽玛函数即函数 $\ln \boldsymbol{\Gamma}(y)$ 加以计算,其中,$\boldsymbol{\Gamma}(\cdot)$ 表示伽玛函数。

第三,多重共线性可能是一个问题。在单指标模型(参见 5.2.4 节)中,要继续对多重共线性实施通常检验。可打印出回归元的相关矩阵,尽管这只考察了两两相关的情况。一种更好的方式是,使用 $\mathbf{X}'\mathbf{X}$ 的条件数,即 $\mathbf{X}'\mathbf{X}$ 的最大特征值与最小特征值之比的平方根。当这个值大于 100 时,就出现问题。对于比单指标模型更为复杂的非线性模型来说,即使条件数并不大,但仍可能有问题。假如人们怀疑多重共性导致了数值问题,就要查看对含有部分变量的模型进行估计是否可行,而这部分变量可能不是共线性的。

第四,在迭代期间不可逆的海赛矩阵并不一定蕴含在真实最大值处出现奇异性。尝试一系列迭代法是值得的,不仅包括牛顿—拉夫森方法,还包含线搜索的最速下降法以及 DFP。该问题还可能起因于多重共线性。

第五,尝试各种不同的初始值。迭代梯度法被设计成获得局部最大值而不是全局最大值。防止出现此类情况的一种方法是,以广泛初始值开始迭代。另一种方法是,实施格点搜索。若 $\boldsymbol{\theta}$ 的维数很大,这两种方法在理论上都需要在许多不同点上进行计算,但是,对于仅包含几个回归元的模型简化形式,详细分析就足够了,尽管模型中的几个回归元大致是统计显著的。

第六,模型可能是不可识别的。实际上,模型识别的标准必要条件是,海赛矩阵是可逆的。如同线性模型一样,简单的检查包括避免虚拟变量陷阱,而且如果在最初分析中使用一部分数据(数据子集),那么在一部分数据中确定所有变量就具有某种变异。例如,如果数据从性别或年龄或地区来看是有序的,那么当这些作为指示变量出现时,就产生了问题,并且所选择的子集是具有特定性别、年龄或地区的个体。对于非线性模型来说,在理论上,很难确定模型是不可识别的。在回到对模型识别进行仔细分析之前,人们经常首先剔除所有其他的潜在原因。

甚至在成功地获得参数估计值之后,仍然会出现计算问题,因为不可能获得方差矩阵 $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}'^{-1}$ 的估计值。当使用迭代法时,比如 DFP,就会产生这种情况,不用海赛矩阵 $\mathbf{A}^{-1}$ 作为迭代中的加权矩阵。首先,例如要检查的是,迭代法确实是收敛的而不是停留在默认的最大迭代次数上。当出现收敛时,就尝试 $\mathbf{A}$ 的一种可供选择的估计,这里通过期望海赛矩阵,或者通过解析导数而不是数值导数,利用更准确的数值来计算。假如这类求解仍然失败,一种可能情况是,模型是不可识别的,

那么在参数估计阶段,对这种不可识别可利用不用计算海赛矩阵的迭代法策略。

人们发现,另一个计算问题是,参数与方差估计并不与先验信息相符合。对于参数估计来说,一些明显检查,包括确保对截距项的正确处理(依赖于内容来决定是包含还是排除)是否达到收敛,以及是否获得全局最大值(通过尝试一系列初始值)。对于不同统计软件包来说,如果参数估计的标准误差都给出相同参数估计,那么最有可能的原因是,各种不同方法都可用于建立方差矩阵估计(参见 5.5.2 节)。

一种好的计算策略是,以一小部分数据及回归元开始,比如说一个回归元与 100 个观测值。倘若项目仅此一个而已,这就可以通过诸如打印出重要输出或者利用嵌入式追踪工具(**trace facility**)进行简化。如果该项目通过检查,那么就整个模型及数据而言,计算问题不可能归因于不正确的数据输入或编码错误,而一种最可能的原因在于真正的计算困难,诸如多重共线性或不好的初始值。

检验项目有效性的一种好方法是,建立模拟数据集,其真实参数是已知的。对于大样本量来说,比如说 $N=10\,000$,估计参数值应接近于真实值。

最后,注意到,从非线性模型估计中所获得的合情合理的计算结果,并不能保证是正确的结果。例如,很多早期出版的多项式 probit 模型的应用明显地报告出敏感性结果,可是被估计的模型结果被确定是不可识别的(参见 15.8.1 节)。

10.5 文献注释

甚至在线性模型中仍会出现数值问题,建议读者阅读戴维森和麦金农(Davidson and MacKinnon, 1993,第 1.5 节)以及格林(Greene, 2003,附录 E)的书。统计计算的标准参考文献是,肯尼迪和金特尔(Kennedy and Gentle,1980),尤其是普雷斯等人(Press et al., 1993)以及普雷斯参与合作的那些书目。对于计算函数来说,标准参考文献是阿布拉莫维茨和斯特根(Abramowitz and Stegun, 1971)。匡特(Quandt, 1983)阐述了许多计算问题,包括最优化。

5.3 迭代法的概述曾由雨宫(Amemiya, 1985,4.4 节)、戴维森和麦金农(Davidson and MacKinnon, 1993,6.7 节)、马达拉(Maddala, 1977,第 9.8 节)特别是格林(Greene, 2003,附录 E6)给出。哈维(Harvey, 1990)曾经给出 GN 算法的许多应用,由于 GN 算法具有简单性,故它是 NLS 估计方面的一种通常迭代法。对于 EM 算法,特别参见雨宫(Amemiya, 1985, 第 375~378 页)。对于 SA,参见戈夫等人的文献(Goffe et al., 1994)。

习 题

10-1 考察当唯一回归元是截距时,logit 回归模型的 MLE 计算。于是,$E[y]=1/(1+e^{-\beta})$,并且已标度的对数似然函数梯度 $g(\beta)=(y-1/(1+e^{-\beta}))$。假定由样本得到 $\bar{y}=0.8$,且初始值是 $\beta=0.0$。

(a) 计算 β 的牛顿—拉夫森算法的前 6 次迭代。

(b) 设式(10.1)中的 $A_s=1$,计算其梯度算法的前 6 次迭代,因而 $\hat{\beta}_{s+1}=\hat{\beta}_s+g_s$。

(c) 比较(a)部分与(b)部分的方法效果。

10-2 考察非线性回归模型 $y=\alpha x_1+\gamma/(x_2-\delta)+u$,其中,$x_1$ 与 x_2 都是外生回归元,它们与 iid 误差 $u\sim\mathcal{N}[0,\sigma^2]$是独立的。

(a) 推导估计(α,γ,δ)的高斯—牛顿算法方程。

(b) 推导估计(α,γ,δ)的牛顿—拉夫森算法方程。

(c) 解释不能任意选取算法初始值的重要性。

10-3 假定 y 的 pdf 具有 C 个成分的混合形式,$f(y|\pi)=\sum_{j=1}^{C}\pi_j f_j(y)$,其中,$\pi=(\pi_1,\cdots,\pi_C)$,$\pi_j>0$,$\sum_{j=1}^{C}\pi_j=1$。$\pi_j$ 是未知的混合比例,而密度 $f_j(y)$的参数是预先假定已知的。

(a) 给定 y_i 的一个随机样本,$i=1,\cdots,N$,写出一般对数似然函数,并求 $\hat{\pi}_{ML}$ 的一阶条件。证明 $\hat{\pi}_{ML}$不存在显式解。

(b) 设 $\mathbf{z}_i$ 表示潜分类变量的$C\times1$ 维向量,$i=1,\cdots,N$,使得当 y 来自混合的第j 个分量时,$z_{ji}=1$,而在其他情况下,$z_{ji}=0$。倘若潜分类变量是可观测的,根据观测到的潜变量写出似然函数。

(c) 推导估计 π 的 EM 算法。(提示:如果 z_{ji}是可观测的,那么 $\hat{\pi}_j$ 的 MLE$=N^{-1}\sum_i z_{ji}$。E 步骤需要计算 $\mathrm{E}[z_{ji}|y_i]$;M 步骤需要用 $\mathrm{E}[z_{ji}|y_i]$代替 z_{ji},然后求解 π)。

10-4 设(y_{1i},y_{2i})服从二元变量正态分布,$i=1,\cdots,N$,其均值为(μ_1,μ_2),且协方差参数$(\sigma_{11},\sigma_{12},\sigma_{22})$,而相关系数为 ρ。假定 y_1 的所有 N 个观测值都可利用,但 y_2 却有 $m<N$ 个缺失观测值。利用 y_i 的边缘分布服从$\mathcal{N}[\mu_j,\sigma_{jj}]$且 $y_2|y_1\sim\mathcal{N}[\mu_{2.1},\sigma_{22.1}]$,其中,$\mu_{2.1}=\mu_2+\sigma_{12}/\sigma_{22}(y_1-\mu_1)$,$\sigma_{22.1}=(1-\rho^2)\sigma_{22}$,请推导估算 y_2[1] 缺失观测值的 EM 算法。

[1] 原著中这里为 y_1,实际应为 y_2。——译者注

第三部分
基于模拟的方法

第二部分已经强调，微观经济计量模型往往对非线性模型进行估计，其数据是从受限于各种抽样偏倚的繁复调查中所抽取的大量且异质的数据。在这种背景下，对经济现象的现实描绘经常需要使用估计及后续统计推断都很困难的模型。现在，计算机硬件和软件的进步使得完成这样的任务成为可能。第三部分阐述现代密集计算和基于模拟的估计与推断方法，此类方法可减少某些困难。处理这种材料的背景会随章节不同而有些变化，但其基础性根基是最小二乘法与极大似然估计。

第 11 章阐述统计推断的自助法。当源自渐近理论的公式很复杂时，这些方法因提供获得标准误差的方法简单而引人注目，例如，如同某些两步估计量的情况。而且，如果实施恰当，自助法能导致更精炼的渐近理论，从而得到小样本较好的统计推断。

第 12 章阐述基于模拟的估计方法。由于不存在得出闭形式解的概率分布的积分，在标准计算方法无法计算估计量的情况下，这些方法使得估计成为可能。

第 13 章概述贝叶斯方法，该方法提供完全不同于本书其他章节所使用的经典方法的估计和推断方法。尽管这是一种不同的方法，但在实际应用中，在大样本背景下贝叶斯方法会产生与那些经典方法相类似的结果。此外，贝叶斯方法在计算形式上以更有效的方式实施。

11 自助法

11.1 引　论

对于大部分微观经济计量学的估计量及其有关的检验统计量来说，不大可能有精确的有限样本结果可以利用。前面几章曾经阐述的统计推断方法，均依赖于通常导致有限正态分布与卡方分布的渐近理论。

一种可供选择的近似是由自助法提供的，该方法归因于埃弗龙(Efron, 1978, 1982)。这通过蒙特卡罗模拟来逼近统计量的分布，其抽样来自经验分布或者观测数据的拟合分布。由于计算能力不断进步，其所需的额外计算通常是可行的。不过，与传统方法一样，自助法依赖于渐近理论，而且仅在无限大样本下是精确的。

将广泛的自助法分成两大类方法。第一类是最简单的自助法，它使得当传统方法诸如标准误差计算很难实施时去进行统计推断。第二类是更复杂的自助法，该方法具有提供可产生有限样本中较好近似的渐近精练的其他优点。应用研究者经常对第一类自助法较感兴趣。而理论学家则强调第二类自助法，尤其是当渐近理论在有限样本条件下表现不好时。

经济计量学文献中关注假设检验对自助法的使用，这依赖于对统计量分布尾部概率的近似。其他一些应用涉及置信区间、标准误差的估计以及缩减偏倚。尽管对含有渐近精练的自助法利用不足，但对于建立在 iid 样本上的光滑$\sqrt{N}$一致估计量来说，可直接实施自助法。在另一些背景下，包括非光滑估计量诸如中位数、非参数估计量以及数据不是 iid 的推断，则要小心谨慎。

11.2 节将对自助法提供一个相当充分的概述，11.3 节给出一个例子，11.4 节则提供某种理论。自助法的进一步变形在 11.5 节加以阐述。11.6 节阐述微观经济计量学中在特定数据形式与特定方法下对自助法的运用。

11.2　自助法概述

我们概述建立在 iid 样本$\{\mathbf{w}_1,\cdots,\mathbf{w}_N\}$基础上的估计量 $\hat{\boldsymbol{\theta}}$ 与有关统计量的重要自助法，这里通常有 $\mathbf{w}_i=(y_i,\mathbf{x}_i)$，而 $\hat{\boldsymbol{\theta}}$ 表示光滑估计量，该估计量是$\sqrt{N}$一致的且服从渐近正态分布。为使记号简单，通常阐述纯量 θ 的一些结果。对于向量 $\boldsymbol{\theta}$，在

大部分例子中，用 θ 的第 j 个分量 θ_j 代替 $\boldsymbol{\theta}$。

关注的统计量包括通常的回归输出：估计值 $\hat{\theta}$；标准误差 $s_{\hat{\theta}}$；t 统计量 $t=(\hat{\theta}-\theta_0)/s_{\hat{\theta}}$，其中，$\theta_0$ 表示零假设值；与这个统计量有关的临界值或 p 值；以及置信区间。

本节将对这些统计量逐一阐述其自助法。而且给出某种动机，而 11.4 节将对基本理论加以概述。

11.2.1 *不带精炼的自助法*

考察样本均值 $\hat{\mu}=\bar{y}=N^{-1}\sum_{i=1}^{N}y_i$ 的方差估计，其中，纯量随机变量 y_i 服从 iid $[\mu,\sigma^2]$，当 $\mathrm{V}[\hat{\mu}]=\sigma^2/N$ 是未知的时候。

$\hat{\mu}$ 的方差能通过取自总体的 S 个容量为 N 的样本而得到，并得出 S 个样本均值；从而得到 S 个估计值 $\hat{\mu}_s=\bar{y}_s$，$s=1,\cdots,S$。然后，我们通过 $(S-1)^{-1}\sum_{s=1}^{S}(\hat{\mu}_s-\bar{\hat{\mu}})^2$ 估计 $\mathrm{V}[\hat{\mu}]$，其中，$\bar{\hat{\mu}}=S^{-1}\sum_{s=1}^{S}\hat{\mu}_s$。

当然，由于我们只有一个样本，所以这一方法不可行。自助法通过把样本看成总体而实施这个方法。于是，有限总体现在就是真实数据 $y_1,\cdots,y_N$。$\hat{\mu}$ 的分布可通过从容量为 N 的此总体中抽取 B 个自助法样本而得到，其中每一个自助法容量为 N 的样本都是通过从 $y_1,\cdots,y_N$ 中进行放回抽样取得的。这就得到 B 个样本均值，从而得出 B 个估计值 $\hat{\mu}_b=\bar{y}_b$，$b=1,\cdots,B$。然后，通过 $(B-1)^{-1}\sum_{b=1}^{B}(\hat{\mu}_b-\bar{\hat{\mu}})^2$ 估计 $\mathrm{V}[\hat{\mu}]$，其中，$\bar{\hat{\mu}}=B^{-1}\sum_{b=1}^{B}\hat{\mu}_b$。放回抽样看起来似乎违背了通常抽样方法，但实际上标准的抽样理论都假定进行放回抽样，而不是不放回抽样(参见 24.2.2 节)。

一旦拥有额外信息，使用其他一些获取自助法样本的方法是可能的。例如，倘若知道 $y_i\sim\mathcal{N}[\mu,\sigma^2]$，我们就能从 $\mathcal{N}[\hat{\mu},s^2]$ 分布中抽取容量为 N 的 B 个自助法样本。这种自助法是参数自助法的一个例子，而前面的自助法则出自经验分布。

更一般地讲，对于估计量 $\hat{\boldsymbol{\theta}}$ 来说，能使用类似的自助法，例如，当 $\mathrm{V}[\hat{\boldsymbol{\theta}}]$ 的解析公式很复杂时，要估计 $\mathrm{V}[\hat{\boldsymbol{\theta}}]$，从而估计标准误差。当观测值 $\mathbf{w}_i$ 对不同 i 是 iid 的时候，这种自助法通常是有效的，而且它们具有类似于运用通常渐近理论所获得的估计性质。

11.2.2 *渐近精炼*

在一些背景下，对前面的自助法加以改进是可能的，并获得等价于运用更加精炼渐近理论所得到的那些估计值，而精炼渐近理论可以更好地逼近 $\hat{\boldsymbol{\theta}}$ 的有限样本分布。本章的大部分内容是针对这类渐近精炼(**asymptotic refinements**)的。

通常，渐近理论使用 $\sqrt{N}(\hat{\theta}-\theta_0)\xrightarrow{d}\mathcal{N}[0,\sigma^2]$ 的结果。因而：

$$\Pr[\sqrt{N}(\hat{\theta}-\theta_0)/\sigma\leqslant z]=\Phi(z)+R_1 \tag{11.1}$$

其中，$\Phi(\cdot)$ 表示标准正态 cdf，而 R_1 表示余项，当 $N\to\infty$ 时，R_1 将消失。

这个结果建立在 5.3 节曾经详述的渐近理论基础之上，包括中心极限定理的应用。CLT(中心极限定理)是建立在截尾幂级数展开式的基础上。11.4.3 节将

详述埃奇沃思展开式，该展开式包括了另外一些项。当具有一个附加项时，得到：

$$\Pr[\sqrt{N}(\hat{\theta}-\theta_0)/\sigma\leqslant z]=\Phi(z)+\frac{g_1(z)\phi(z)}{\sqrt{N}}+R_2 \tag{11.2}$$

其中，$\phi(\cdot)$表示标准正态密度，$g_1(\cdot)$表示给定 11.4.3 节的式(11.13)之后的有界函数，而 R_2 表示余项，当 $N\to\infty$时，该余项将会消失。

埃奇沃思展开式在理论很难实施，因为函数 $g_1(\cdot)$以复杂方式成为数据相依的。含有渐近精炼(**with asymptotic refinement**)的自助法提供了一种简单的计算方法实施埃奇沃思展开式。该理论将在 11.4.4 节给出。

由于 $R_1=O(N^{-1/2})$且 $R_2=O(N^{-1})$，所以在渐近形式上 $R_2<R_1$，当 $N\to\infty$时会产生更好的近似。不过，在有限样本中，可能出现 $R_2>R_1$。含有渐近精炼的自助法在渐近形式上提供一种更好的近似，这种近似导致希望典型使用的有限容量样本会更好地近似。然而，不存在这样的保证，而且模拟研究经常用于验证有限样本确实存在好处。

11.2.3 *渐近中枢统计量*

为了出现渐近精炼，作为自助法的统计量必须是渐近中枢统计量(**asymptotically pivotal statistic**)，这意味着该统计量的极限分布不依赖于未知参数。该结果将在 11.4.4 节加以解释。

举一个例子，考察从 $y_i\sim[\mu,\sigma^2]$中进行抽样。于是，估计值 $\hat{\mu}=\bar{y}\overset{a}{\sim}\mathcal{N}[0,\sigma^2/N]$就不是渐近中枢的，甚至给定零假设值 $\mu=\mu_0$ 时，因为它的分布依赖于未知参数 σ^2。然而，学生化统计量(**studentized statistic**) $t=(\hat{\mu}-\mu_0)/s_{\hat{\mu}}\overset{a}{\sim}\mathcal{N}[0,1]$则是渐近中枢的。

一般地讲，估计量不是渐近中枢的。然而，常规的渐近标准正态或卡方分布检验统计量，包括沃尔德、拉格朗日乘子、似然比检验以及有关的置信区间都是渐近中枢的。

11.2.4 *自助法*

本节我们将对自助法给出一种广泛的描述，更进一步的详细内容，将在后面各节展开。

自助法算法

一般自助法算法(**bootstrap algorithm**)如下：

1. 已知数据 $\mathbf{w}_1,\cdots,\mathbf{w}_N$，利用下面将给出的方法抽取容量为 N 的自助样本，并将新样本记为 $\mathbf{w}_1^*,\cdots,\mathbf{w}_N^*$。

2. 利用自助样本计算适当的统计量。一些例子包括：(a) θ 的估计值$\hat{\theta}^*$；(b) 估计值 $\hat{\theta}^*$ 的标准误差 $s_{\hat{\theta}^*}$；(c) 在最初估计值 $\hat{\theta}$ 处中心化的 t 统计量 $t^*=(\hat{\theta}^*-\hat{\theta})/s_{\hat{\theta}^*}$。这里，$\hat{\theta}^*$ 与 $s_{\hat{\theta}^*}$ 均是以通常方式使用新自助样本而不是最初样本计算的。

3. 对步骤 1 与 2 各自重复 B 次，其中，B 表示很大的数，得到关注统计量的 B 次自助复制，比如 $\hat{\theta}_1^*,\cdots,\hat{\theta}_B^*$ 或 $t_1^*,\cdots,t_B^*$。

4. 运用 B 次自助复制，获得统计量的自助法形式，正如下面一小节所详述的。

具体实施会依据自助样本怎样获得、执行多少次自助法、什么样的统计量作为自助的以及该项统计量是否是渐近中枢的而变化。

自助法抽样法

在步骤 1 中的自助法数据生成过程(dgp)用于逼近真实的未知数据生成过程。

最简单的自助方法是使用数据的经验分布，将样本看成是总体。然后，通过从 $\mathbf{w}_1, \cdots, \mathbf{w}_N$ 中进行放回抽样获得 $\mathbf{w}_1^*, \cdots, \mathbf{w}_N^*$。这样做得到的每个自助样本中，最初数据的某些点将多次出现，而另外一些点将根本不出现。这一方法是经验分布函数(EPF)自助法[**empirical distribution function (EDF) bootstrap**]或非参数自助法(**nonparametric bootstrap**)。它也被称为成对自助法(**paired bootstrap**)，因为在单方程回归模型中有 $\mathbf{w}_i = (y_i, \mathbf{x}_i)$，所以这里既对 y_i 再抽样，又对 $\mathbf{x}_i$ 再抽样。

假定对数据的条件分布进行设定，比如说 $y|\mathbf{x} \sim F(\mathbf{x}, \boldsymbol{\theta}_0)$，同时可利用估计值 $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$。然后，尽管通过从 $F(\mathbf{x}_i, \hat{\boldsymbol{\theta}})$ 中随机抽取来生成 y_i，但我们在步骤 1 中运用最初的 $\mathbf{x}_i$ 来形成自助样本。这对应于重复样本中固定回归元(参见 4.4.5 节)。否则，我们首先从 $\mathbf{x}_1, \cdots, \mathbf{x}_N$ 中重新抽取 $\mathbf{x}_i^*$，然后从 $F(\mathbf{x}_i^*, \hat{\boldsymbol{\theta}})$ 中生成 $y_i, i=1, \cdots, N$。这两个都是能应用于完全参数模型的参数自助法(**parametric bootstrap**)的例子。

对于含有可加 iid 误差的回归模型来说，比如说 $y_i = g(\mathbf{x}_i, \boldsymbol{\beta}) + u_i$，我们能形成拟合残差 $\hat{u}_1, \cdots, \hat{u}_N$，其中，$\hat{u}_i = y_i - g(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$。然后，在步骤 1 中，从这些残差中自助而得到一个新的残差取样，比如说$(\hat{u}_1^*, \cdots, \hat{u}_N^*)$，从而得到自助样本$(y_1^*, \mathbf{x}_1), \cdots, (y_N^*, \mathbf{x}_N)$，其中 $y_i^* = g(\mathbf{x}_i, \hat{\boldsymbol{\beta}}) + u_i^*$。这种自助法称为残差自助法(**residual bootstrap**)。它使用了介于非参数自助法与参数自助法之间的信息。如果误差项具有不依赖于未知参数的分布，那么就能应用残差自助法。

我们之所以强调成对自助法，是因为它对广泛的非线性模型具有简单性和可应用性。然而，其他一些自助法通常会提供更好的近似[参见霍罗维茨(Horowitz, 2001, 第 3 185 页)]，并且如果它们需要的较强模型假设都得以成立，就应该加以应用。

自助法的次数

自助法特性依赖于 $N \to \infty$，因此甚至对于较小的 B，自助法在渐近形式上都是有效的。不过，很明显当 $B \to \infty$时，自助法更为准确。充分大的 B 会随着导入自助法(**bootstrap-induced**)模拟误差的容许与自助法的目的而变化。

安德鲁斯和布基斯基(Andrews and Buchinsky, 2000)已经阐述了确保给定精度水平，或者等价地对于给定 B 值所获得的精度水平，用特定应用的数值方法决定所需要的复制次数 B。设 λ 表示关注的一个量，譬如标准误差或临界值，$\hat{\lambda}_\infty$ 表示满足 $B=\infty$ 的理想自助法估计值，而 $\hat{\lambda}_B$ 表示具有 B 次自助法的估计值。然后，安德鲁斯和布基斯基证明：

$$\sqrt{B}(\hat{\lambda}_B - \hat{\lambda}_\infty)/\hat{\lambda}_\infty \xrightarrow{d} \mathcal{N}[0, \omega]$$

其中，ω 随着应用而变，并由安德鲁斯和布基斯基的表 III 加以定义。由此可得，$\Pr[\delta \leqslant z_{\tau/2}\sqrt{\omega/B}] = 1-\tau$，其中，$\delta = |\hat{\lambda}_B - \hat{\lambda}_\infty|/\hat{\lambda}_\infty$ 表示仅仅由 B 次复制所引起的相

对偏差。因而，$B \geqslant \omega z_{\tau/2}^2/\delta^2$ 确保了相对偏差至少以概率 $1-\tau$ 小于 δ。否则，给定 B 次复制，相对偏差小于 $\delta = z_{\tau/2}\sqrt{\omega/B}$。

为了提供具体的指导原则，我们提出下述经验法则：

$$B = 384\omega$$

该式确保了相对偏差至少以概率 0.95 小于 10%，因为 $z_{0.025}^2/0.1^2=384$。实施中唯一的困难部分是对 ω 的估计，这将随应用而变化。

对于标准误差估计来说，$\omega=(2+\gamma_4)/4$，其中，γ_4 表示自助法估计量 $\hat{\theta}^*$ 的超峰度的系数。从直观上讲，估计量分布的较肥尾部(**fatter tail**)意味着可能有异常值，污染了标准误差估计。由此可得，当 $\gamma_4=0$ 时，$B=384\times(1/2)=192$ 就足够，而当 $\gamma_4=8$ 时，需要 $B=960$。这些值均大于由埃弗龙和蒂布沙兰尼(Efron and Tibsharani, 1993, 第 52 页)曾经提出的那些值，他们认为，$B=200$ 几乎总是足够的。

对于对称双侧检验或者水平为 α 的置信区间来说，$\omega=\alpha(1-\alpha)/[2z_{\alpha/2}\phi(z_{\alpha/2})]^2$。当 $\alpha=0.05$ 时，得到 $B=348$，而当 $\alpha=0.01$ 时，得到 $B=685$。如人们所料，要进一步深入研究分布尾部，就需要更多次自助法。

对于单侧检验或者非对称双侧检验或水平为 α 的置信区间来说，$\omega=\alpha(1-\alpha)/[z_\alpha\phi(z_\alpha)]^2$。当 $\alpha=0.05$ 时，得到 $B=634$，而当 $\alpha=0.01$ 时，得到 $B=989$。当对尾部进行检验时，需要更多次的自助法。对于自由度为 h 的卡方检验来说，$\omega=\alpha(1-\alpha)/[\chi_\alpha^2(h)f(\chi_\alpha^2(h))]^2$，其中，$f(\cdot)$表示 $\chi_\alpha^2(h)$)密度。

对于检验 p 值来说，$\omega=(1-p)/p$。例如，当 $p=0.05$ 时，$\omega=19$，而且 $B=7\,296$。与超过临界值的拒绝假设相比，要准确计算检验 p 值就需要更多次自助法。

对于 θ 的偏倚修正估计来说，简单规则使用 $\hat{\omega}=\hat{\sigma}^2/\hat{\theta}^2$，其中，估计量 $\hat{\theta}$ 具有标准误差 $\hat{\sigma}$。例如，如果通常 t 统计量 $t=\hat{\theta}/\hat{\sigma}=2$，那么 $\hat{\omega}=1/4$，而 $B=96$。安德鲁斯和布基斯基(Andrews and Buchinsky, 2000)曾提供许多更详细的内容和对这些结果的细致改进。

对于假设检验来说，戴维森和麦金农 (Davidson and MacKinnon, 2000)提供了一种可供选择的方法。他们关注于由具有有限 B 次自助法引起的功效损失。(注意到，如果 $B=\infty$，就没有损失功效。)在模拟基础上，他们对于水平为 0.05 的检验，建议至少 $B=399$，而对于水平为 0.01 的检验至少 $B=1\,499$。他们论证了，其检验方法优于安德鲁斯和布基斯基的那种方法。

麦金农(MacKinnon, 2002)曾经概述了由戴维森和麦金农撰写的其他几篇论文内容，强调自助法推断中出现的一些实践问题。对于水平为 α 的假设检验来说，选取 B 以使 $\alpha(B+1)$为整数。例如，当 $\alpha=0.05$ 时，设 $B=399$ 而不是 400。相反，如果 $B=400$，那么对上面单侧而言，第 20 个或第 21 个最大自助法 t 统计量是否是临界值就不清楚了。对于非线性模型来说，可以在每个自助样本中令初值等于初始参数估计值，仅仅执行几次牛顿—拉夫森迭代而简化计算。

11.2.5 *标准误差估计*

估计量的方差自助法估计(**bootstrap estimate of variance**)是将通常的估计方

差公式应用于 B 次自助法复制 $\hat{\theta}_1^*,\cdots,\hat{\theta}_B^*$ 上：

$$s_{\hat{\theta},\mathrm{Boot}}^2=\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b^*-\bar{\hat{\theta}}^*)^2 \tag{11.3}$$

其中：

$$\bar{\hat{\theta}}^*=B^{-1}\sum_{b=1}^{B}\hat{\theta}_b^* \tag{11.4}$$

取平方根得到 $s_{\hat{\theta},\mathrm{Boot}}$，即标准误差的自助法估计值(**bootstrap estimate of the standard error**)。

这个自助法并没有提供渐近精炼。然而，当利用传统方法很难获得标准误差时，它却有令人惊奇的用途。存在许多这样的例子。估计值 $\hat{\theta}$ 可以是序贯两步 m 估计量[1](**sequential two-step m-estimator**)，其标准误差利用由 6.8 节给出的结果很难计算出。估计值 $\hat{\theta}$ 是利用软件包所估计出的 2SLS 估计量，而该软件包只报告假定同方差时的标准误差，但误差实际上是异方差的(**heteroskedastic**)。估计值 $\hat{\theta}$ 可以是实际所估计的其他参数函数(**function of other parameters**)，例如 $\hat{\theta}=\hat{\alpha}/\hat{\beta}$，而使用自助法来代替 δ 方法。对于含有许多小整群的整群数据[2](**clustered data**)来说，诸如短面板，整群稳健的标准误差可通过从整群中再抽样而获得。

由于自助法估计 $s_{\hat{\theta},\mathrm{Boot}}$ 是一致的，它可用于代替通常渐近公式中的 $s_{\hat{\theta}}$ 来构成渐近有效的置信区间与假设检验。因而，在很难通过其他方法获得标准误差的背景下，渐近统计推断就是可行的。然而，有限样本实施将不存在改进(**no improvement**)。为了获得渐近精炼，需要下一节的一些方法。

11.2.6 *假设检验*

这里，我们考察个体系数的检验，系数记为 θ。检验或者是单侧向上的选择 $H_0:\theta\leqslant\theta_0$ 与 $H_a:\theta>\theta_0$，或者是双侧检验 $H_0:\theta=\theta_0$ 与 $H_a:\theta\neq\theta_0$。其他一些检验推迟到 11.6.3 节。

含有渐近精炼的检验

通常检验统计量 $T_N=(\hat{\theta}-\theta_0)/s_{\hat{\theta}}$ 提供了渐近精炼的潜力，这是因为它的渐近标准正态分布不依赖于未知参数，从而成为渐近中枢的。我们执行 B 次自助法复制生成 B 个检验统计量 $t_1^*,\cdots,t_B^*$，其中：

$$t_b^*=(\hat{\theta}_b^*-\hat{\theta})/s_{\hat{\theta}_b^*} \tag{11.5}$$

估计值 t_b^* 集中围绕在初始估计值 $\hat{\theta}$ 的附近，因为再抽样是从集中于 $\hat{\theta}$ 附近的分布中进行抽取的。经验分布 $t_1^*,\cdots,t_B^*$，从小到大排列，用于逼近 T_N 分布。

对于单侧向上的可选择检验来说，自助法临界值(**bootstrap critical value**)(在 α 水平上)是 B 个有序检验统计量的向上 α 分位数。例如，当 $B=999$ 且 $\alpha=0.05$ 时，临界值是 t^* 的第 950 个最大值，从而 $(B+1)(1-\alpha)=950$。对于单侧向下的可供

[1] 又称为序列两步 m 估计量。——译者注
[2] 又称为群聚数据。——译者注

选择检验，临界值是 t^* 的第 50 个最小值。

人们也可以明显方式计算自助法 p 值（**bootstrap p-value**）。例如，当最初统计量 t 位于 999 个自助法复制的第 914 个与第 915 个最大值之间，单侧向上的可选择检验的 p 值是 $1-914/(B+1)=0.086$。

对于双侧检验来说，需要在对称检验与非对称检验之间加以区别。对于非对称（**nonsymmetrical test**）或等尾部检验（**equal-tailed test**）而言，自助法临界值（**critical values**）（在 α 水平上）是有序统计量 t^* 的向上 $\alpha/2$ 与向下 $\alpha/2$ 的分位数，而且当原始 t 统计量位于这个范围之外，就拒绝零假设。然而，对于对称检验（**symmetrical test**）而言，我们对 $|t^*|$ 排序，其自助法临界值（在 α 水平上）是有序 $|t^*|$ 的向上 α 分位数。当 $|t|$ 大于这个范围时，就在 α 水平上拒绝零假设。

利用 t 百分位数方法（**percentile-t method**）的这些检验，提供了渐近精炼。对于单侧 t 检验与非对称的双侧 t 检验，检验的真实水平（**true size**）是带有标准渐近临界值的 $\alpha+O(N^{-1/2})$ 与带有自助法临界值的 $\alpha+O(N^{-1})$。对于双侧对称 t 检验或者渐近卡方检验，渐近近似会实施得更好，并且，利用标准渐近临界值进行检验的真实水平是 $\alpha+O(N^{-1})$，而利用自助临界值进行检验的真实水平是 $\alpha+O(N^{-2})$。

不含渐近精炼的检验

尽管渐近有效并没有提供渐近精炼，但可使用一种可供选择的自助法。

一种曾在 11.2.5 节末尾提及的方法是计算 $t=(\hat{\theta}-\theta_0)/s_{\hat{\theta},\text{Boot}}$，其中，由式（11.3）给出的自助估计值 $s_{\hat{\theta},\text{Boot}}$ 代替通常的估计值 $s_{\hat{\theta}}$，同时把这个检验统计量与出自标准正态分布的临界值相比。

第二种方法，此处阐述双侧检验 $H_0: \theta=\theta_0$ 与 $H_a: \theta\neq\theta_0$ 是求出自助法估计值 $\hat{\theta}_1^*, \cdots, \hat{\theta}_B^*$ $\alpha/2$ 的向下 $\alpha/2$ 分位数与向上 $\alpha/2$ 分位数，而当 θ_0 落入这一范围之外，就拒绝 H_0。这称为百分位数方法（**percentile method**）。通过利用以 $\hat{\theta}$ 为中心而不是以 θ_0 为中心的式（11.5）中 t_b^*，并在每一步自助法中利用不同的标准误差 $s_{\hat{\theta}}^*$，可以获得渐近精炼。

这两种自助法的优点是，不需要计算 $s_{\hat{\theta}}$，其中，$s_{\hat{\theta}}$ 是建立在渐近理论基础上的通常标准误差估计值。

11.2.7 置信区间

绝大多数统计学著作都考察置信区间估计，而将假设检验放置在一边。相反，我们在这里以假设检验开始，不过有必要给出置信区间的一个简略表述。

渐近精炼建立在 t 统计量基础之上，是渐近中枢的。因此，由 11.2.4 节中的步骤 1～3，我们可获得自助复制的 t 统计量 $t_1^*, \cdots, t_B^*$。于是，设 $t^*_{[1-\alpha/2]}$ 与 $t^*_{[\alpha/2]}$ 表示这些 t 统计量的向下 $\alpha/2$ 与向上 $\alpha/2$ 分位数。t 百分位数方法（**percentile-t method**）的 $100(1-\alpha)$ 百分数置信区间是：

$$(\hat{\theta}+t^*_{[1-\alpha/2]}\times s_{\hat{\theta}},\ \hat{\theta}+t^*_{[\alpha/2]}\times s_{\hat{\theta}})^{[1]} \tag{11.6}$$

[1] 原著中开区间内左边数为 $\hat{\theta}-t^*_{[1-\alpha/2]}\times s_{\hat{\theta}}$，应为 $\hat{\theta}+t^*_{[1-\alpha/2]}\times s_{\hat{\theta}}$。——译者注

其中，$\hat{\theta}$ 与 $s_{\hat{\theta}}$ 分别是来自最初样本的估计值与标准误差。

一种可供选择的方法是埃弗龙(Efron，1987)曾详述的偏倚校正与加速(BC_a)方法[**bias-corrected and accelerated (BC_a) method**]。这在比 t 百分位数方法更广泛的问题类型上提出了渐近精炼。

虽然其他一些方法提供渐近有效置信区间，但不带渐近精炼。首先，人们能够使用通常置信区间公式中标准误差的自助法估计值，得到区间($\hat{\theta}-z_{[1-\alpha/2]}\times s_{\hat{\theta},\mathrm{Boot}}$，$\hat{\theta}+z_{[\alpha/2]}\times s_{\hat{\theta},\mathrm{Boot}}$)。其次，百分位数方法(**percentile method**)置信区间是 θ 的 B 次自助法估计值 $\hat{\theta}_1^*,\cdots,\hat{\theta}_B^*$ 的向下 $\alpha/2$ 与向上 $\alpha/2$ 的分位数的距离。

11.2.8 *偏倚缩减*

通常，非线性估计量在有限样本中是有偏的，尽管如果这个估计量是一致的，那么偏倚会渐近地趋于0。例如，当 μ^3 通过 $\hat{\theta}=\bar{y}^3$ 来估计时，$\mathrm{E}[\hat{\theta}-\mu^3]=3\mu\sigma^2/N+\mathrm{E}[(y-\mu)^3]/N^2$，其中，$y_i$ 服从 iid $[\mu,\sigma^2]$。

更一般地讲，对于 $\sqrt{N}$ 一致估计量来说，有：

$$\mathrm{E}[\hat{\theta}-\theta_0]=\frac{a_N}{N}+\frac{b_N}{N^2}+\frac{c_N}{N^3}+\cdots \tag{11.7}$$

其中，a_N、b_N 以及 c_N 都表示有界常值，只是这些常值会随着数据和估计量而变化[参见霍尔(Hall，1992，第53页)]。一个可供选择的估计量 $\tilde{\theta}$ 提供了渐近精炼，如果：

$$\mathrm{E}[\tilde{\theta}-\theta_0]=\frac{B_N}{N^2}+\frac{C_N}{N^3}+\cdots \tag{11.8}$$

其中，B_N 与 C_N 都表示有界常值。当 $N\to\infty$ 时，这两个估计量的偏倚均会消失。后者因其偏倚以较快的速率趋于0而引人注目，因此它是渐近精炼的，尽管在有限样本中，可能有 $(B_N/N^2)>(a_N/N+b_N/N^2)$。

我们想要估计偏倚 $\mathrm{E}[\hat{\theta}]-\theta$。这是参数的期望值或总体平均值与参数所生成的数据之间的距离。由于自助法代替带有样本的总体，所以自助样本是通过参数 $\hat{\theta}$ 生成的，它具有关于自助法的平均值 $\bar{\hat{\theta}}^*$。于是，此偏倚自助法估计值(**bootstrap estimate of the bias**)是：

$$\mathrm{Bias}_{\hat{\theta}}=(\bar{\hat{\theta}}^*-\hat{\theta}) \tag{11.9}$$

其中，$\bar{\hat{\theta}}^*$ 已在式(11.4)中定义。

例如，假定 $\hat{\theta}=4$，且 $\bar{\hat{\theta}}^*=5$。于是，估计的偏倚是 $(5-4)=1$，即向上偏倚1。一旦给定偏倚修正估计为3，由于 $\hat{\theta}$ 被过高估计到1，所以偏倚校正需要从 $\hat{\theta}$ 中减去1。更一般地讲，θ 的自助偏倚修正估计值(**bootstrap bias-corrected estimator**)是：

$$\begin{aligned}\hat{\theta}_{\mathrm{Boot}}&=\hat{\theta}-(\bar{\hat{\theta}}^*-\hat{\theta})\\&=2\hat{\theta}-\bar{\hat{\theta}}^*\end{aligned} \tag{11.10}$$

注意到，$\bar{\hat{\theta}}^*$ 自身并不是偏倚修正估计值。关于校正方向看起来令人困惑，更详细内容，参见埃弗龙和蒂布沙兰尼(Efron and Tibsharani，1993，第138页)。对于典型

的$\sqrt{N}$一致估计量来说，$\hat{\theta}$ 的渐近偏倚是 $O(N^{-1})$，而 $\hat{\theta}_{\text{Boot}}$ 的渐近偏倚是 $O(N^{-2})$。

在实际应用中，对于$\sqrt{N}$一致估计量，偏倚校正几乎很少使用，因为与最初估计值 $\hat{\theta}$ 相比，自助法估计值变化更大，而且其偏倚常常相对小于估计值的标准误差。自助法偏倚校正用于收敛速率小于$\sqrt{N}$的那些估计量，尤其是非参数回归与密度估计量。

11.3 自助法例子

举一个自助法例子，考察在 5.9 节引入的指数回归模型。这里的数据是由指数分布生成的，该指数分布的指数均值具有两个回归元：

$$y_i \mid \mathbf{x}_i \sim \text{指数}(\lambda_i), \quad i=1,\cdots,50$$

$$\lambda_i = \exp(\beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i})$$

$$(x_{2i}, x_{3i}) \sim \mathcal{N}[0.1,\ 0.1;\ 0.1^2,\ 0.1^2,\ 0.005]$$

$$(\beta_1,\ \beta_2,\ \beta_3) = (-2,\ 2,\ 2)$$

对 50 个观测值的样本进行极大似然估计，得到 $\hat{\beta}_1 = -2.192$，$\hat{\beta}_2 = 0.267$，$s_2 = 1.417$，而 $t_2 = 0.188$；并且 $\hat{\beta}_3 = 4.466$，$s_3 = 1.741$ 而 $t_3 = 2.679$。就这个 ML 例子而言，标准误差建立在 $-\hat{\mathbf{A}}^{-1}$ 即负的估计海赛矩阵逆之上。

我们集中关注对 β_3 的统计推断，并且阐明标准误差计算、统计显著性检验、置信区间以及偏倚校正方面的自助法。自助法估计与通常的渐近估计之间的差异在本例中相对较小，但在其他例子中却可能相当大。

此处报告的一些结果都是根据以放回方式对$(y_i,\ x_{2i},\ x_{3i})$联合再抽样 $B=999$ 次的成对自助法。由表 11.1 知，999 自助复制估计值 $\hat{\beta}^*_{3,b}$，具有均值 4.716 且标准差 1.939，$b=1,\cdots,999$。表示 11.1 还给出 $\hat{\beta}^*_3$ 与 t^*_3（下面将要定义）的重要百分位数。

表 11.1　关于斜率系数的自助法统计推断：例子[a]

	$\hat{\beta}^*_3$	t^*_3	$z=t(\infty)$	$t(47)$
均值	4.716	0.026	1.021	1.000
SD[b]	1.939	1.047	1.000	1.021
1%	−0.336	−2.664	−2.326	−2.408
2.5%	0.501	−2.183	−1.960	−2.012
5%	1.545	−1.728	−1.645	−1.678
25%	3.570	−0.621	−0.675	−0.680
50%	4.772	0.062	0.000	0.000
75%	5.971	0.703	0.675	0.680
95%	7.811	1.706	1.645	1.678
97.5%	8.484	2.066	1.960	2.012
99.0%	9.427	2.529	2.326	2.408

[a] 概括统计量与百分位数都建立在 999 对自助法再抽样之上，关于：(1) 估计值 $\hat{\beta}^*_3$；(2) 有关统计量 $t^*_3 = (\hat{\beta}^*_3 - \hat{\beta}_3)/s_{\hat{\beta}^*_3}$；(3) 具有 47 个自由度的学生 t 分布；(4) 标准正态分布。最初，数据生成过程是从正文给出的指数分布中所抽取的，样本量为 50。

[b] SD 表示标准差。

然而，能使用参数自助法。于是，自助法样本通过从具有参数 $\exp(\hat{\beta}_1+\hat{\beta}_2x_{2i}+\hat{\beta}_3x_{3i})$ 的指数分布中抽取 y_i 而获得。不过，在对 H_0：$\beta_3=0$ 进行检验的情况下，该指数参数可以是 $\exp(\tilde{\beta}_1+\tilde{\beta}_2x_{2i})$，其中，$\tilde{\beta}_1$ 与 $\tilde{\beta}_2$ 都是来自最初样本的约束极大似然估计值。

标准误差：由式(11.3)，标准误差的自助法估计值可对 β_3 的 999 次自助复制估计利用通常的标准差公式来计算。与通常的渐近标准误差估计值 1.741 相比，这会得到估计值 1.939。注意到，这一自助法没有提供精炼而仅仅用作核对，或者如果通过其他手段被证明很难求出标准误差。

含有渐近精炼的假设检验：我们考察在水平 0.05 上对 H_0：$\beta_3=0$ 与 H_a：$\beta_3\neq0$ 进行检验。含有渐近精炼的检验建立在 t 统计量之上，是渐近中枢的。由 11.2.6 节知，对于每个自助法，我们都可以计算 $t_3^*=(\hat{\beta}_3^*-4.664)/s_{\hat{\beta}_3^*}$，它是以来自最初样本的 $\hat{\theta}_3=4.664$ 为中心的。对于非对称检验来说，自助法临界值等于 t_3^* 的 999 个值中向上与向下的 2.5 百分位数，即第 25 个最小值与第 25 个最大值。由表 11.1 知，这些值分别是 -2.813 与 2.066。由于来自最初样本计算出的 t 统计量 $t_3=(4.466-0)/1.741=2.679>2.066$，所以拒绝零假设。不过，使用了 $|t_3^*|$ 的向上 5 百分位数的对称检验，得到自助法临界值 2.078，这再次导致在水平 0.05 上拒绝 H_0。

在这个例子中，自助法临界值大于那些利用标准正态的或 $t(47)$ 的渐近近似的临界值，并且为这一目的而安排的有限样本被在正态性下线性模型的准确结果所激发而进行调整。因此，此例子中通常的渐近结果导致了过度拒绝，同时所具有的实际水平大于名义水平。例如，在 5%上，z 区域临界值(-1.960，1960)小于其自助法临界值(-2.183，2066)。图 11.1 画出建立在利用核方法滑光化的 t 检验密度 t_3^* 之上的自助法估计值，并将它与标准正态情况相对比。这两个密度看起来很接近，尽管其左边尾部显著地比自助法估计要宽一些。

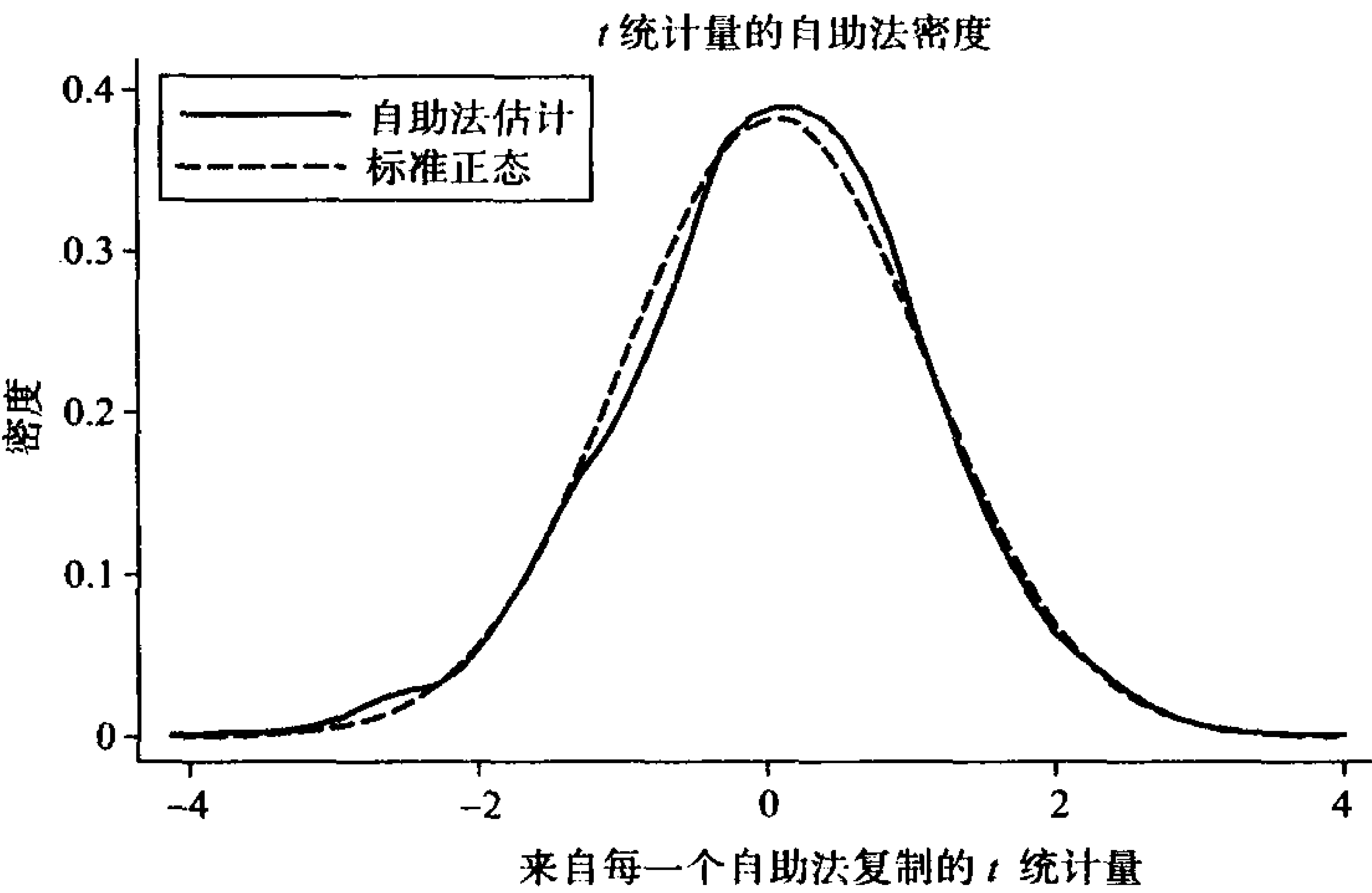


图 11.1 关于斜率等于 0 的 t 检验统计量的自助法密度可从 999 次自助复制中获得，将其与画出的具有标准正态密度情况相对比。数据从指数分布回归模型中生成。

不含渐近精炼的假设检验：可以使用一种可供选择的自助法检验方法，却没有施加渐近精炼。首先，一旦利用自助发标准误差估计值 1.939，而不是渐近标准误差估计值 1.741，得到 $t_3=(4.664-0)/1.939=2.405$。这就导致在水平 0.05 上拒绝，或者利用标准正态的临界值，或者利用 $t(47)$ 的临界值。其次，由表 11.1 知，自助法估计值 $\hat{\beta}_3^*$ 的 95%位于(0.501, 8.484)之中，并没有包括所假定的值 0，因而，我们再次拒绝 $H_0:\beta_3=0$。

置信区间：渐近精炼可通过利用 95%百分位数 t 置信区间来获得。当应用式(11.6)时，得到$(4.664-2.183\times1.741,\ 4.664+2.066\times1.741)$或者(0.864, 8.260)。这与传统的 95%渐近置信区间$(4.664-1.960\times1.741, 4.664+1.960\times1.741)$或者(1.25, 8.08)相比较。

能建立其他一些置信区间，只是这些区间没有渐近精炼。一旦利用自助法标准误差估计，得到 95%置信区间为$(0.864,\ 8.464)=4.664\pm1.960\times1.939$。百分位数方法使用 999 自助法系数估计的向上 2.5 与向下 2.5 百分位数，得出 95%置信区间 (0.501, 8.484)。

偏倚校正：与最初估计值 4.664 相比，β_3 的 999 自助复制估计的均值是 4.716。特别地，与 $s_3=1.741$ 的标准误差相比，估计偏倚 $(4.716-4.664)=0.052$。估计偏倚是向上的，而且由式(11.10)得到，β_3 的偏倚修正估计等于 $4.664-0.052=4.612$。

自助法依赖于渐近理论，并且实际上可以提供有限样本近似，与传统方法相比要差一些。为了证实自助法实是一种改进，这里我们需要完全蒙特卡罗分析，比如说从具有指数数据的生成过程中抽取一个容量为 50 的 1 000 个样本，然后对这些样本的每一个进行自助法，比如说进行 999 次。

11.4 自助法理论

本节解释遵循霍罗维茨(Horowitz, 2001)的全面概述。一些重要结果是关于自助法一致性的，而且如果自助法应用于渐近中枢统计量上，那就是渐近精炼的。

11.4.1 *自助法*

我们用 $X_1,\cdots,X_N$ 作为数据的一般记号，其中，为了记号简单，不采用黑体 X_i，即使通常它是一个向量，例如 $(y_i,\mathbf{x}_i)$。假定数据是从具有 cdf $F_0(x)=\Pr[X\leqslant x]$的分布中独立抽取的。在最简单的应用中，F_0 是有限维数族，满足 $F_0=F_0(x,\boldsymbol{\theta}_0)$。

将所考察的统计量记为 $T_N=T_N(X_1,\cdots,X_N)$。T_N 的准确有限样本分布是 $G_N=G_N(t,F_0)=\Pr[T_N\leqslant t]$。一个问题是求出对 G_N 的良好近似。

传统渐近理论使用 T_N 的渐近分布，记为 $G_\infty=G_\infty(t,F_0)$。这在理论上可能依赖于未知 F，在此情况下，我们利用 F_0 的一致估计。例如，使用 $\hat{F}_0=F_0(\cdot,\hat{\boldsymbol{\theta}})$，其中，$\hat{\boldsymbol{\theta}}$ 关于 $\boldsymbol{\theta}_0$ 是一致的。

经验自助法采取一种截然不同的方法逼近 $G_N(\cdot,F_0)$。不是用 G_∞ 代替 G_N，而

是用 F_0 的一致估计量 F_N，比如样本的经验分布代替总体 cdf F_0。

虽然在解析形式上不能确定 $G_N(\cdot, F_N)$，但可通过自助法来逼近它。一种含有放回自助法的再抽样会得到统计量 $T_N^* = T_N(X_1^*, \cdots, X_N^*)$。独立重复这一步骤 B 次，得出复制 $T_{N,1}^*, \cdots, T_{N,B}^*$。$T_{N,1}^*, \cdots, T_{N,B}^*$ 的经验 cdf 是 T 分布的自助法估计，得到：

$$\hat{G}_N(t, F_N) = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}(T_{N,b}^* \leqslant t) \tag{11.11}$$

其中，当事件 A 发生时，$\mathbf{1}(A)$ 等于 1，否则 $\mathbf{1}(A)$ 等于 0。这正是关于 $T_N^* \leqslant t$ 已实现的自助法再抽样的比例。

记号已总结在表 11.2 中。

表 11.2　自助法理论记号

数　量	记　号
样本 (iid)	$X_1, \cdots, X_N$，其中，X_i 通常表示向量
X 的总体 cdf	$F_0 = F_0(x, \boldsymbol{\theta}_0) = \Pr[X \leqslant x]$
关注的统计量	$T_N = T_N(X_1, \cdots, X_N)$
T_N 的有限样本 cdf	$G_N = G_N(t, F_0) = \Pr[T_N \leqslant t]$
T_N 的极限 cdf	$G_\infty = G_\infty(t, F_0)$
T_N 的渐近 cdf	$\hat{G}_\infty = G_\infty(t, \hat{F}_0)$，其中，$\hat{F}_0 = F_0(x, \hat{\boldsymbol{\theta}})$
T_N 的自助法 cdf	$\hat{G}_N(t, F_N) = B^{-1}\sum_{b=1}^{B} \mathbf{1}(T_{N,b}^* \leqslant t)$

11.4.2 *自助法一致性*

很明显，当自助法次数 $B \to \infty$ 时，自助法估计值 $\hat{G}_N(t, F_N)$ 收敛于 $G_N(t, F_N)$。因此，自助法估计值 $\hat{G}_N(t, F_N)$ 关于 $G_N(t, F_0)$ 的一致性需要：

$$G_N(t, F_N) \xrightarrow{p} G_N(t, F_0)$$

这里，关于统计量是一致的，并且是对于使 cdf 存在的空间中的所有 F_0。

显然，F_N 关于 F_0 必是一致的。另外，需要关于 dgp $F_0(x)$ 的光滑性（**smoothness**），因此，对于很大的 N，在一些观测值 x 上，$F_N(x)$ 与 $F_0(x)$ 一致地互相接近。此外，需要关于 $G_N(\cdot, F)$ 的光滑性，即所考察统计量的 cdf 作为 F 的函数，因而，当 N 很大时，$G_N(\cdot, F_N)$ 接近 $G_N(\cdot, F_0)$。

霍罗维茨（Horowitz, 2001，第 3 166～3 168 页）给出两个正式定理，一个是一般性的，另一个则是关于 iid 数据的，同时提供自助法潜在失效的一些例子，包括中位数的估计以及具有界参数约束的估计。

受限于 F_N 关于 F_0 的一致性以及需要 F_0 与 G_N 的光滑性，自助法会产生一致估计与渐近有效推断。自助法在相当广泛的背景下是一致的。

11.4.3 *埃奇沃思展开式*

自助法的另一个引人注目之处是，它考虑到了渐近精炼。辛格（Singh, 1981）

曾经提供利用埃奇沃思展开式的证明，现在我们就加以介绍。

考察 $Z_N=\sum_i X_i/\sqrt{N}$的渐近特性，其中，为了简单起见，X_i 表示标准化的纯量随机变量，它服从 iid [0, 1]。然后，应用中心极限定律，得到 Z_N 的极限标准正态分布。更准确地讲，Z_N具有 cdf：

$$G_N(z)=\Pr[Z_N\leqslant z]=\Phi(z)+O(N^{-1/2}) \tag{11.12}$$

其中，$\Phi(\cdot)$表示标准正态 cdf。余项可被忽略，并且常规的渐近理论可通过 $G_\infty(z)=\Phi(z)$来逼近 $G_N(z)$。

利用中心极限定律可推导式(11.12)，这可通过对 Z_N 特征函数(**characteristic function**) $\mathrm{E}[e^{isZ_N}]$的简单近似来正式推导，其中 $i=\sqrt{-1}$ 。一种较好的近似可使这个特征函数以 $N^{-1/2}$幂形式展开。通常的埃奇沃思展开式(**Edgeworth expansion**)增加两项，从而有：

$$G_N(z)=\Pr[Z_N\leqslant z]=\Phi(z)+\frac{g_1(z)}{\sqrt{N}}+\frac{g_2(z)}{N}+O(N^{-3/2}) \tag{11.13}$$

其中，$g_1(z)=-(z^2-1)\phi(z)\ \kappa_3/6$，$\phi(\cdot)$表示标准正态密度，$\kappa_3$ 表示 Z_N 的第三个累积量，而关于 $g_2(\cdot)$的很长的表达式是由罗滕伯格(Rothenberg, 1984, 第 895 页)或雨宫(Amemiya, 1985，第 93 页)给出。一般地，第 r 个半不变量[1](**cumulant**)κ_r 是对数特征函数或累积量母函数的 $\ln(\mathrm{E}[e^{isZ_N}])=\sum_{r=0}^{\infty}\kappa_r(is)^r/r!$级数展开式中的第 r 个系数。

式(11.13)中的余项可以被忽略，而埃奇沃思展开式可通过 $G_\infty(z, F_0)=\Phi(z)+N^{-1/2}g_1(z)+N^{-1}g_2(z)$ 来逼近 $G_N(z, F_0)$。如果 Z_N 是一个检验统计量，那么这能用于计算 p 值与临界值。否则，对式(11.13)求逆：

$$\Pr\left[Z_N+\frac{h_1(z)}{\sqrt{N}}+\frac{h_2(z)}{N}\leqslant z\right]\simeq\boldsymbol{\Phi}(z) \tag{11.14}$$

其中，函数 $h_1(z)$与 $h_2(z)$已由罗滕伯格(Rothenberg, 1984，第 895 页)给出。其左边给出一个修正统计量，它通过标准正态的 Z_N 而不是最初统计量 Z_N 得到较好的近似。

应用中出现的问题是，Z_N 的半不变量需要计算一些函数 $g_1(z)$与 $g_2(z)$或者 $h_1(z)$与 $h_2(z)$。对于这些半不变量来说，很难获得其解析表达式[例如，萨根(Sargan, 1980)；还有菲利普斯(Phillips, 1983)]。自助法提供一种数值方法来实施不需要计算半不变量的埃奇沃思展开式，正如下面所证明的。

11.4.4 *渐近精炼与自助法*

现在，回到 11.4.1 节更一般的背景上，额外假设是 T_N 具有极限分布且可应用通常的$\sqrt{N}$渐近特性。

一些传统的渐近方法，使用极限 cdf $G_\infty(t,F_0)$作为对真实 cdf $G_N(t,F_0)$的近

[1] 又称为累积量。——译者注

似。对于$\sqrt{N}$一致渐近正态估计量来说，这具有如下误差，该误差极限特性拥有$N^{-1/2}$的倍数形式。我们将其写成：

$$G_N(t,F_0)=G_\infty(t,F_0)+O(N^{-1/2}) \tag{11.15}$$

其中，$G_\infty(t,F_0)=\Phi(t)$。

一种较好的近似是，可能利用埃奇沃思展开式。于是：

$$G_N(t,F_0)=G_\infty(t,F_0)+\frac{g_1(t,F_0)}{\sqrt{N}}+\frac{g_2(t,F_0)}{N}+O(N^{-3/2}) \tag{11.16}$$

不幸的是，正如已注意到的，右边函数$g_1(\cdot)$与$g_2(\cdot)$很难构造。现在，考察自助法估计量$G_N(t,F_N)$。由埃奇沃思展开式，得到：

$$G_N(t,F_N)=G_\infty(t,F_N)+\frac{g_1(t,F_N)}{\sqrt{N}}+\frac{g_2(t,F_N)}{N}+O(N^{-3/2}) \tag{11.17}$$

详细内容参见霍尔(Hall，1992)。自助法估计量$G_N(t,F_N)$可用于逼近有限样本的 cdf $G_N(t,F_0)$。当用式(11.17)减去式(11.16)，得到：

$$\begin{aligned}G_N(t,F_N)-G_N(t,F_0)=&[G_\infty(t,F_N)-G_\infty(t,F_0)]\\&+\frac{[g_1(t,F_N)-g_1(t,F_0)]}{\sqrt{N}}+O(N^{-1})\end{aligned} \tag{11.18}$$

假定F_N关于真实 cdf F_0是一致的，所以$F_N-F_0=O(N^{-1/2})$。对于连续函数G_∞来说，式(11.18)右边第一项$[G_\infty(t,F_N)-G_\infty(t,F_0)]$是$O(N^{-1/2})$，因而$G_N(t,F_N)-G_N(t,F_0)=O(N^{-1/2})$。

因此，自助法近似$G_N(t,F_N)$一般并不比通常渐近近似$G_\infty(t,F_0)$更渐近地接近于$G_N(t,F_0)$，参见式(11.15)。

现在，假定统计量T_N是渐近中枢的(**asymptotically pivotal**)，因此它的渐近分布G_∞不依赖于未知参数。此处正是下述情况：如果T_N是标准化的，那么它的极限分布是正态分布。于是，$G_\infty(t,F_N)=G_\infty(t,F_0)$，因而式(11.18)简化成：

$$G_N(t,F_N)-G_N(t,F_0)=N^{-1/2}[g_1(t,F_N)-g_1(t,F_0)]+O(N^{-1}) \tag{11.19}$$

然而，由于$F_N-F_0=O(N^{-1/2})$，所以对于F中连续的g_1来说，我们有$[g_1(t,F_N)-g_1(t,F_0)]=O(N^{-1/2})$。由简化结果可得，$G_N(t,F_N)=G_N(t,F_0)+O(N^{-1})$。现在，自助法近似$G_N(t,F_N)$是$G_N(t,F_0)$的一个较好渐近近似，因为其误差现在是$O(N^{-1})$。

总之，就渐近中枢统计量上的自助法而言，我们有：

$$G_N(t,F_0)=G_N(t,F_N)+O(N^{-1}) \tag{11.20}$$

它是对传统近似$G_N(t,F_0)=G_\infty(t,F_0)+O(N^{-1/2})$的一种改进。

因此，在下述意义下，基于渐近中枢统计量的自助法会导致一种改进的小样本表现。设α是检验程序的名义水平。通常的渐近理论会产生具有真实水平$\alpha+O(N^{-1/2})$的t检验，而自助法会产生具有真实水平$\alpha+O(N^{-1})$的t检验。

对于对称双侧假设检验与置信区间来说，可以证明，与利用通常渐近理论所产

生的误差 $O(N^{-1})$相比，在渐近中枢统计量之上的自助法具有近似误差 $O(N^{-3/2})$。

前面结果被限制在渐近正态统计量上。对于卡方分布检验统计量来说，其渐近好处类似于那些对称双侧假设检验。借助于自助法来证明偏倚缩减，参见霍罗维茨(Horowitz, 2001，第 3172 页)。

理论分析会产生下述要点。自助法应形成关于 F_0 一致的分布 F_N。自助法要求关于 F_0 与 G_N 的光滑性以及连续性，因此这时需要对标准自助法加以修正，例如，因为参数边界约束诸如 $\theta \geqslant 0$，所以存在不连续。自助法假定低阶矩存在，因为低阶累积量出现在埃奇沃思展开式的函数 g_1 之中。渐近精炼需要使用渐近中枢统计量。所阐述的自助法精炼均假定 iid 数据，因此甚至需要修改异方差误差。对于更完整的讨论，参见霍罗维茨(Horowitz, 2001)。

11.4.5 *自助法检验功效*

自助法分析关注于小样本具有正确水平的检验。如同任何水平校正一样，自助法的水平校正将导致检验功效的变化。

从直观上讲，当利用一阶渐近检验的真实水平大于名义水平，具有渐近精炼的自助法不仅减少名义水平大小，因为拒绝会很少发生，而且也减少了检验功效。反之，当真实水平小于名义水平，自助法将增大检验功效。在霍罗维茨(Horowitz, 1994，第 409 页)的模拟应用中，观察到了这一情况。有意思的是，在他的模拟研究中发现，尽管渐近等价于检验的自助一阶会产生具有类似真实水平(基本上等于名义水平)的检验，但在不同的自助法检验中，检验功效存在相当大的差异。

11.5 自助法推广

至今，所述的自助法强调基于 iid 数据的光滑的$\sqrt{N}$一致渐近正态估计量。下述对自助法的推广允许在更广泛范围内应用一致自助法(11.5.1 节与 11.5.2 节)或含有渐近精炼的一致自助法(11.5.3 节～11.5.5 节)。对于这些更高等的方法，只进行简略阐述。某些方法将在 11.6 节运用。

11.5.1 *二次抽样方法*

二次抽样方法(**subsampling method**)使用的样本量 m 本质上比样本量 N 更小一些。二次抽样可以是放回的[比克尔、戈策和范 · 茨韦特(Bickel, Gotze, and van Zwet, 1997)]，也可以是不放回的[波利特斯和罗马诺(Politis and Romano, 1994)]。

放回二次抽样提供作为总体的随机样本的子样本，而不是对分布估计的随机样本，诸如在成对自助法(**paired bootstrap**)情况下的样本。于是，当 11.4.2 节已经讨论的光滑性条件失效产生完全样本自助法的不一致性时，放回二次抽样可以是一致的。不过，有关检验或置信区间的渐近误差却具有比使用不带精炼的完全样本自助法所获得的通常 $O(N^{-1/2})$更高阶的量。

当完全样本自助法无效的时候，子样本是有益的，或者作为验证完全样本自助

法为有效的一种方法。其结果将随子样本量的选择而有所不同。此外，由于使用样本较小部分，所以会相当大地增加样本误差。实际上，我们应该具有$(m/N)\to 0$且$N\to\infty$。波利特斯、罗马诺和沃尔夫(Politis, Romano, and Wolf, 1999)以及霍罗维茨(Horowitz, 2001)都提供更进一步的详情。

11.5.2 *移动分块自助法*

移动分块自助法(**moving blocks bootstrap**)用于数据相关的而不是独立的情况。这要将样本分割成r个非重叠的、长度为l的块(**blocks**)，其中，$rl\simeq N$。首先，人们从这些块中进行放回抽样，得到r个新的分块，这会具有不同于原来r个块的临时排序。然后，人们利用这种自助法样本估计参数。移动分块方法是，将随机抽取的块作为每一个都互相独立，只允许块内出现相关。安德森(Anderson, 1971)实际上使用过类似分块来推导m个相依过程的中心极限定理。移动分块过程要求，当$N\to\infty$时$r\to\infty$，以确保我们可能推导相邻组每一个都不相关。还要求，当$N\to\infty$时，分块长度$l\to\infty$。例如，参见约策和孔施(Götze and Künsch, 1996)。

11.5.3 *嵌入式自助法*

由霍尔(Hall, 1986)、贝兰(Beran, 1987)与隆(Lon, 1987)引进的嵌套式自助法(**nested bootstrap**)是自助法之中套自助法。当自助法不是建立在渐近中枢的统计量之上时，这一方法尤其有用。特别地，若很难计算估计值的标准误差，则人们可对当前自助法样本运用自助法来获得自助法标准误差估计值$s_{\hat{\theta}^*,\text{自助法}}$，并构成$t^*=(\hat{\theta}^*-\hat{\theta})/s_{\hat{\theta}^*,\text{自助法}}$，然后对自助法复制$t_1^*,\cdots,t_B^*$运用百分位数$t$方法。这使渐近精炼成为可能，而单一自助法是无法实现的。

更一般地讲，迭代自助法(**iterated bootstrapping**)是一种通过估计源自经过一次自助法的误差(也就是偏倚)并且修正这些误差而改进自助法效果的方法。通常，如果统计量是渐近中枢的，那么自助法的每一次进一步迭代都会减少偏倚系数N^{-1}，否则减少偏倚系数$N^{-1/2}$。参见霍尔和马丁(Hall and Martin, 1988)给出的一种良好解释。如果在每一次迭代中都执行B次自助法，那么当存在k次迭代，就要求实施B^k自助法。鉴于此，至多执行两次迭代，称为双迭代(**double bootstrap**)或者标定自助法(**calibrated bootstrap**)。

戴维森、欣克利和谢克特曼(Davison, Hinkley, and Schechtman, 1986)曾经提出平衡自助法(**balanced bootstrapping**)。这种方法保证了每个样本观测值都是准确地重复使用所有B次自助法的相同数目，得出一个更好的自助法估计。有关实施内容，参见格利森(Gleason, 1988)，他的算法与通常非平衡自助法相比，只是增加一点计算时间。

11.5.4 *重新中心化与再标度*

为了获得渐近精炼，自助法应建立在对正考虑的模型施加所有条件的数据生成过程F_0的估计值$\hat{F}_0$之上。一个重要例子是含有残差的自助法。

在非线性模型中，甚至在线性模型中，当没有截距时，最小二乘法残差之和不

为 0。于是，建立在最小二乘法残差之上的残差自助法（参见 11.2.4 节）就利用约束 $\mathrm{E}[u_i]=0$ 而言将失效。相反，残差自助法应该对重新中心化残差（**recentered residual**）$\hat{u}_i-\bar{\hat{u}}$进行自助法，其中，$\bar{\hat{u}}=N^{-1}\sum_{i=1}^{N}\hat{u}_i$。类似地，重新中心化应该在过度识别模型中对 GMM 估计量的序实施自助法（参见 11.6.4 节）。

对残差重新标度（**rescaling**）也是有用的。例如，在含有 iid 误差的线性回归模型中，从$(N/(N-K))^{1/2}\hat{u}_i$ 中再抽样，因为这些都具有方差 s^2。其他一些调整包括利用标准化残差$\hat{u}_i/\sqrt{(1-h_{ii})s^2}$，其中，$h_{ii}$表示射影矩阵 $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$中的第 i 个对角元素。

11.5.5 刀切法

自助法能用于偏倚修正（参见 11.2.8 节）。一种可供选择的再抽样方法是刀切法，即自助法的前身。刀切法使用 N 个规定性定义的样本量为 $N-1$ 的子样本，它们是通过依次去掉 N 个观测值中的每一个而获得的，然后重新计算其估计量。

为了理解刀切法是如何起作用的，设 $\hat{\theta}_N$ 表示利用所有 N 个观测值的 θ 的估计值，同时设 $\hat{\theta}_{N-1}$表示前$(N-1)$个观测值 θ 的估计值。如果式(11.7)成立。那么 $\mathrm{E}[\hat{\theta}_N]=\theta+a_N/N+b_N/N^2+O(N^{-3})$，而且 $\mathrm{E}[\hat{\theta}_{N-1}]=\theta+a_N/(N-1)+b_N/(N-1)^2+O(N^{-3})$，这蕴含 $\mathrm{E}[N\hat{\theta}_N-(N-1)\hat{\theta}_{N-1}]=\theta+O(N^{-2})$。因而，与 $\hat{\theta}_N$ 偏倚相比，$N\hat{\theta}_N-(N-1)\hat{\theta}_{N-1}$的偏倚更小。

可是，此估计量更易变化，因为它使用较少的数据。举一个例子，倘若 $\hat{\theta}=\bar{y}$，新的估计量就是 y_N，即第 N 个观测值。同理，此变异能通过去掉每个观测值且进行平均而得以减少。

于是，更正式地，考察建立在源自 iid 数据的样本量 N 之上的参数向量 $\boldsymbol{\theta}$ 的估计量$\hat{\boldsymbol{\theta}}$ 。对于 $i=1,\cdots,N$，顺次删除第 i 个观测值，进而从 N 个刀切法再抽样容量为$(N-1)$的样本中获得 N 个刀切复制估计值$\hat{\boldsymbol{\theta}}_{(-i)}$。$\hat{\boldsymbol{\theta}}$ 偏倚的刀切法估计值（**jackknife estimate of the bias**）是$(N-1)(\bar{\hat{\boldsymbol{\theta}}}-\hat{\boldsymbol{\theta}})$，其中，$\bar{\hat{\boldsymbol{\theta}}}=N^{-1}\sum_i\hat{\boldsymbol{\theta}}_{(-i)}$表示 N 个刀切法复制$\hat{\boldsymbol{\theta}}_{(-i)}$的平均。由于用$(N-1)$相乘，偏倚看起来似乎很大，但其差$(\hat{\boldsymbol{\theta}}_{(-i)}-\hat{\boldsymbol{\theta}})$却比自助法情况下的小很多，因为刀切法再抽样样本不同于原始样本，仅仅相差一个观测值。

这就产生对 $\boldsymbol{\theta}$ 刀切法估计值（**jackknife estimate**）的偏倚校正：

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\text{刀切法}}&=\hat{\boldsymbol{\theta}}-(N-1)(\bar{\hat{\boldsymbol{\theta}}}-\hat{\boldsymbol{\theta}})\\&=N\hat{\boldsymbol{\theta}}-(N-1)\bar{\hat{\boldsymbol{\theta}}}\end{aligned}\tag{11.21}$$

这使偏倚从 $O(N^{-1})$缩减到 $O(N^{-2})$，这与自助法情况的偏倚缩减是同阶的。至于自助法，假定其估计量是光滑的$\sqrt{N}$一致的估计量。刀切法估计与 $\hat{\boldsymbol{\theta}}$ 相比，具有增大的方差，并且刀切法失效的一些例子已由米勒（Millor，1974）给出。

一个简单的例子是，来自满足 $y_i\sim[\mu,\sigma^2]$ iid 样本对 σ^2 的估计。估计值$\hat{\sigma}^2=N^{-1}\sum_i(y_i-\bar{y})^2$，MLE 在正态性下具有 $\mathrm{E}[\hat{\sigma}^2]=\sigma^2(N-1)/N$，因此，其偏倚等于 σ^2/N，它是一个 $O(N^{-1})$。在此例子中，可以证明，刀切法估计被简化成 $\hat{\sigma}^2_{\text{刀切法}}=$

$(N-1)^{-1}\sum_i(y_i-\bar{y})^2$，人们并不要求计算 N 个独立的估计值 $\hat{\sigma}^2_{(-i)}$。这是 σ^2 的一个无偏估计值，因而该偏倚实际上是 0，而不是通常的 $O(N^{-2})$ 结果。

刀切法归功于克努取(Quenouille, 1956)。图基(Tukey, 1958)考虑了在更广泛统计学中的应用问题。特别地，估计量 $\hat{\theta}$ 的标准误差刀切法估计值是：

$$\widehat{\mathrm{se}}_{刀切法}[\hat{\theta}]=\left[\frac{N-1}{N}\sum_{i=1}^{N}(\hat{\theta}_{(-i)}-\bar{\hat{\theta}})^2\right] \tag{11.22}$$

图基通过仿照可以求解各种问题的多功能武器库(Boy Scout jackknife，又译为童子军大刀)提出了刀切法术语，其中的每一个都通过特殊构造的工具得以更有效地解决。

在许多情形下，刀切法是用于缩减偏倚"粗略但尚能使用"的方法，但它不是任何情况下的理想方法。刀切法能被看成是自助法的一种线性近似[埃弗龙和蒂布沙兰尼(Efron and Tibsharani, 1993，第 146 页)]。在小样本条件下，与自助法相比，它要求较少的计算，从而 $N<B$ 是可能的，却胜过当 $B\to\infty$ 时借助于自助法。

考察线性回归模型 $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{u}$，满足 $\hat{\boldsymbol{\beta}}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$。源自 OLS 回归的偏倚估计量的例子是，含有滞后因变量作为回归元的时间序列模型。基于第 i 个刀切法样本($\mathbf{X}_{(-i)}$, $\mathbf{y}_{(-i)}$)的回归估计量是由

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{(-i)}&=[\mathbf{X}'_{(-i)}\mathbf{X}_{(-i)}]^{-1}\mathbf{X}'_{(-i)}\mathbf{y}_{(-i)}\\&=[\mathbf{X}'\mathbf{X}-\mathbf{x}_i\mathbf{x}_i']^{-1}(\mathbf{X}'\mathbf{y}-\mathbf{x}_iy_i)\\&=\hat{\boldsymbol{\beta}}-[\mathbf{X}'\mathbf{X}]^{-1}\mathbf{x}_i(y_i-\mathbf{x}_i'\hat{\boldsymbol{\beta}}_{(-i)})\end{aligned}$$

给出的。第三个等式避开了对每个 i 需要求 $\mathbf{X}'_{(-i)}\mathbf{X}_{(-i)}$ 逆，而这可利用

$$[\mathbf{X}'\mathbf{X}]^{-1}=[\mathbf{X}'_{(-i)}\mathbf{X}_{(-i)}]^{-1}-\frac{[\mathbf{X}'_{(-i)}\mathbf{X}_{(-i)}]^{-1}\mathbf{x}_i\mathbf{x}_i'[\mathbf{X}'_{(-i)}\mathbf{X}_{(-i)}]^{-1}}{1+\mathbf{x}_i'[\mathbf{X}'_{(-i)}\mathbf{X}_{(-i)}]^{-1}\mathbf{x}_i}$$

求得。此处，伪值(**pseudo-values**)是由 $N\hat{\boldsymbol{\beta}}-(N-1)\hat{\boldsymbol{\beta}}_{(-i)}$给出的，并且 $\hat{\boldsymbol{\beta}}$ 的刀切法估计量由

$$\hat{\boldsymbol{\beta}}_{\mathrm{Jack}}=N\hat{\boldsymbol{\beta}}-(N-1)\frac{1}{N}\sum_{i=1}^{N}\hat{\boldsymbol{\beta}}_{(-i)} \tag{11.23}$$

给出。

刀切法对偏倚缩减的有趣应用是刀切法 IV 估计量(参见 6.4.4 节)。

11.6 自助法应用

我们考察典型考虑到微观经济计量中一些复杂问题的自助法的应用，诸如异方差性、聚集以及能导致简单自助法失效的复杂估计量。

11.6.1 异方差性误差

对于含有异方差性的、可加误差模型的最小二乘法来说，标准方法是使用怀特异方差性一致协方差矩阵估计量(HCCME)。众所周知，对小样本而言，这样做表

现不好。要是做得正确,自助法会提供一种改进。

成对自助法会导致有效推断,因为$(y_i, \mathbf{x}_i)$是 iid 的基本假设还允许 $V[u_i|\mathbf{x}_i]$ 随 $\mathbf{x}_i$ 而变化(参见 4.4.7 节)。不过,它并没有提供一种渐近精炼,因为它没有利用 $E[u_i|\mathbf{x}_i]=0$ 这一条件。

通常的残差自助法确实导致无效推断,由于它假定 $u_i|\mathbf{x}_i$ 为 iid 的,因此错误利用了同方差误差的条件。依据 11.4 节理论,$\hat{F}$ 关于 F 是非一致的。人们能设定异方差性的正式模型,比如说 $u_i=\exp(\mathbf{z}_i'\alpha)\varepsilon_i$,其中,$\varepsilon_i$ 是 iid 的,得到估计值$\exp(\mathbf{z}_i'\hat{\alpha})$,然后对隐含残差 $\hat{\varepsilon}_i$ 进行自助法。这种自助法的一致性与渐近精炼都要求对异方差性函数形式加以正确设定。

原始自助法(**wild bootstrap**)是由吴(Wu, 1986)与刘(Liu, 1988)引进的,而马曼(Mamme,1993)做出进一步研究,提供对异方差性没有利用这类结构的渐近精炼。这种自助法是用下述残差:

$$\hat{u}_i^*=\begin{cases}\dfrac{1-\sqrt{5}}{2}\hat{u}_i\simeq-0.618\,0\,\hat{u}_i, & \text{以概率}\dfrac{1+\sqrt{5}}{2\sqrt{5}}\simeq0.723\,6\\ \left[1-\dfrac{1-\sqrt{5}}{2}\right]\hat{u}_i\simeq1.618\,0\,\hat{u}_i, & \text{以概率 }1-\dfrac{1+\sqrt{5}}{2\sqrt{5}}\simeq0.276\,4\end{cases}$$

仅对两点分布取期望,并经过某些代数运算,得到 $E[\hat{u}_i^*]=0$, $E[\hat{u}_i^{*2}]=\hat{u}_i^2$,而 $E[\hat{u}_i^{*3}]=\hat{u}_i^3$。因而,$\hat{u}_i^*$ 产生了人们希望的零条件均值,因为 $E[\hat{u}_i^*|\hat{u}_i,\mathbf{x}_i]=0$ 蕴含 $E[\hat{u}_i^*|\mathbf{x}_i]=0$,而二阶矩与三阶矩都是不变的。

原始自助法再抽样拥有第 i 个观测值$(y_i^*, \mathbf{x}_i)$,其中 $y_i^*=\mathbf{x}_i'\hat{\boldsymbol{\beta}}+\hat{u}_i^*$。由于$\hat{u}_i^*$的实现值各不相同,所以再抽样会变化。霍罗维茨(Horowitz, 1997, 2001)通过模拟证明,当存在异方差性时,这种自助法与成对自助法相比执行得更有效,甚至不存在异方差性时,也比其他自助法执行得好。

看起来令人惊讶,因为就第 i 个观测值而言,它仅仅从两个可能残差值 $-0.618\,0\hat{u}_i$ 或 $1.618\,0\hat{u}_i$ 中抽取,这种自助法应该起作用。可是,对于所有 N 个观测值与所有 B 次自助法迭代可进行类似的抽取。同样回想起,怀特的估计量是用 $\hat{u}_i^2$ 代替 $E[u_i^2]$,它尽管对一个观测值是不正确的,却对样本平均值而言是有效的。然而,原始自助法从服从均值为 0 且方差为 $\hat{u}_i^2$ 的两点分布中进行抽取。

11.6.2 *面板数据与整群数据*

考察线性面板回归模型:

$$\tilde{y}_{it}=\tilde{\mathbf{w}}_{it}'\boldsymbol{\theta}+\tilde{u}_{it}$$

其中,i 表示个体,而 t 表示时期。遵循 21.2.3 节的记号,例如,添加 ~ 表示原始数据 y_{it},并首先对 $\mathbf{x}_{it}$ 进行变换剔除固定效应。我们假定,误差 $\tilde{u}_{it}$ 对不同 i 是独立的,尽管 $\tilde{u}_{it}$ 可能是异方差的,并且给定 i 时对 t 来说是相关的。

当面板是短的,故 T 是有限的,而且渐近理论依赖于 $N\to\infty$,故 $\hat{\boldsymbol{\theta}}$ 的一致标准误差能通过成对自助法或 EDF 自助法来获得,这里对不同 i 进行再抽样,却不是

对不同 t 进行再抽样。在前面表述中,$\mathbf{w}_i$ 变成$[y_{i1},\mathbf{x}_{i1},\cdots,y_{iT},\mathbf{x}_{iT}]$,从而我们对 i 进行再抽样而获得选定 i 的全部 T 个观测值。

这种面板自助法(**panel bootstrap**)也称为分块自助法(**block bootstrap**),它同样可用于第 23 章的非线性面板模型。其重要假设是,面板是短的且数据对不同的 i 是独立的。更一般地讲,倘若群容量是有限的,并且整群数目趋于无穷大,当数据是整群的(参见 24.5 节),就可以应用这种自助法。

面板自助法可产生渐近等价于面板稳健三明治误差的标准误差(参见 21.2.3 节)。它却不会产生渐近精炼。不过,它执行起来相当简单,同时在实践上相当有用,尽管甚至关于十分基本的面板估计量,诸如固定效应估计量,许多软件包并没有自动地提供面板稳健标准误差。倘若再抽样又一次地仅仅对 i 进行,其他一些自助法,比如参数自助法与残差自助法,可能是可行的,但要依赖于应用而定。

若误差是 iid 的,则渐近精炼就容易做。不过,更现实地讲,$\bar{u}_{it}$ 将是异方差的且对给定 i 时不同 t 是相关的。如果面板是短的,那么线性模型中的原始自助法(参见 11.6.1 节)应提供渐近精炼。然后,原始自助法再抽样具有(i,t)个观测值$(\tilde{y}_{it}^*,\tilde{\mathbf{w}}_{it})$,其中,$\tilde{y}_{it}^*=\tilde{\mathbf{w}}_{it}'\boldsymbol{\theta}+\hat{\tilde{u}}_{it}^*$,$\hat{\tilde{u}}_{it}=\tilde{y}_{it}-\tilde{\mathbf{w}}_{it}'\hat{\boldsymbol{\theta}}$,而 $\hat{u}_{it}^*$ 表示从 11.6.1 节给出的两点分布中所抽取的。

11.6.3 *假设检验与设定检验*

11.2.6 节曾关注对假设 $\theta=\theta_0$ 的检验。这里,我们考察更一般的检验。如同 11.2.6 节一样,自助法可用于执行含有渐近精炼或不含渐近精炼的假设检验。

不含渐近精炼的检验

自助法无效的一个重要例子是豪斯曼检验(参见 8.3 节)。执行这种标准检验需要估计 $\mathrm{V}[\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}}]$,其中,$\hat{\boldsymbol{\theta}}$ 与 $\tilde{\boldsymbol{\theta}}$ 是两个相互比较的估计量。要获得该估计值很困难,除非做出强假设:两个估计量之一在 H_0 下是完全有效的。不过,运用成对自助法,得到一致估计值:

$$\hat{\mathrm{V}}_{\mathrm{Boot}}[\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}}]=\frac{1}{B-1}\sum_{b=1}^{B}[(\hat{\boldsymbol{\theta}}_b^*-\tilde{\boldsymbol{\theta}}_b^*)-(\bar{\hat{\boldsymbol{\theta}}}^*-\bar{\tilde{\boldsymbol{\theta}}}^*)][(\hat{\boldsymbol{\theta}}_b^*-\tilde{\boldsymbol{\theta}}_b^*)-(\bar{\hat{\boldsymbol{\theta}}}^*-\bar{\tilde{\boldsymbol{\theta}}}^*)]'$$

其中,$\bar{\hat{\boldsymbol{\theta}}}^*=B^{-1}\sum_b\hat{\boldsymbol{\theta}}_b^*$,而 $\bar{\tilde{\boldsymbol{\theta}}}^*=B^{-1}\sum_b\tilde{\boldsymbol{\theta}}_b^*$。然后,计算:

$$\mathrm{H}=(\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}})'(\hat{\mathrm{V}}_{\text{自助法}}[\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}}])^{-1}(\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}}) \tag{11.24}$$

同时与卡方临界值比较。正如第 8 章提及的,需要使用广义逆,并且要小心谨慎,确保利用正确自由度来获得卡方临界值。

更一般地讲,这一方法可用于任何执行起来很困难的标准正态检验或卡方分布检验,因为必须要估计方差。一些例子包括,基于两步估计量的假设检验以及第 8 章的 m 检验。

含有渐近精炼的检验

许多检验尤其是那些完全参数模型,比如 LM 检验与 IM 检验,都能利用辅助回归来直接进行(参见 7.3.5 节和 8.2.2 节)。不过,作为结果的检验统计量在有限样板中执行欠佳,正如许多蒙特卡罗研究所证实的。这类检验统计量很容易计

算，并且是渐近中枢的，因为卡方分布并不依赖于未知参数。原因在于它们是通过自助法进行渐近精炼的重要备选者。

考察 H_0 的 m 检验：$\mathrm{E}[\mathbf{m}_i(y_i|\mathbf{x}_i,\boldsymbol{\theta})]=\mathbf{0}$ 与 H_a：$\mathrm{E}[\mathbf{m}_i(y_i|\mathbf{x}_i,\boldsymbol{\theta})]\neq\mathbf{0}$(参见 8.2 节)。由最初数据，通过 ML 估计 $\hat{\boldsymbol{\theta}}$，并计算检验统计量 M。一旦利用参数自助法，从拟合条件密度 $f(y_i|\mathbf{x}_i,\hat{\boldsymbol{\theta}})$ 中再抽样 y_i^*，对于重复样本中的固定回归元来说，或者从 $f(y_i|\mathbf{x}_i^*,\hat{\boldsymbol{\theta}})$ 中进行再抽样 y_i^*。计算自助法再抽样样本中的 M_b^*，$b=1,\cdots,B$。当最初计算的统计量 M 大于 M_b^* 的 α 分位数时，在水平 α 上就拒绝 H_0，$b=1,\cdots,B$。

霍罗维茨(Horowitz, 1994)已经阐述过 IM 检验的这种自助法，并利用关于这种自助法的坚实的有限样本好处例子来加以证明。德鲁克(Drukker, 2002)对 Tobit 模型给出了设定检验的一个详细应用，提出条件矩设定检验很容易地应用于完全参数模型，因为辅助回归中的任何水平扭曲(**size distortion**)能通过自助法加以修正。注意到，不含渐近精炼的自助法检验，诸如此处给出的豪斯曼检验，可借助于 11.5.3 节给出的嵌入式自助法加以精炼。

11.6.4 *过度识别 GMM、最小距离与经验似然*

GMM 估计量是建立在总体矩条件 $\mathrm{E}[\mathbf{h}(\mathbf{w}_i,\boldsymbol{\theta})]=\mathbf{0}$ 之上的(参见 6.3.1 节)。在恰好识别模型中，一致估计量可直接求解 $N^{-1}\sum_i\mathbf{h}(\mathbf{w}_i,\hat{\boldsymbol{\theta}})]=\mathbf{0}$。在过度识别模型中，这种估计量不再可行。相反，却可使用 GMM 估计量(参见 6.3.2 节)。

现在，考察利用成对自助法或 EDF 自助法来进行自助。对于过度识别模型中的 GMM 来说，$N^{-1}\sum_i\mathbf{h}(\mathbf{w}_i,\hat{\boldsymbol{\theta}})\neq\mathbf{0}$，所以此自助法没有对自助法再抽样样本施加最初的总体约束 $\mathrm{E}[\mathbf{h}(\mathbf{w}_i,\boldsymbol{\theta})]=\mathbf{0}$。因此，即使可以使用渐近中枢 t 统计量，也不存在自助法精炼，但 $\hat{\boldsymbol{\theta}}$ 的自助法与有关的置信区间以及 t 检验统计量仍是一致的。更为基本的是，可以证明，OIR 检验的自助法(参见 6.3.8 节)是非一致的。我们虽然关注于横截面数据，但过度识别模型中的面板 GMM 估计量(参见第 22 章)却会产生类似的问题。

霍尔和霍罗维茨(Hall and Horowitz, 1996)提供了通过重新中心化(**recentering**)来对此加以修正。于是，自助法就建立在 $\mathbf{h}^*(\mathbf{w}_i,\hat{\boldsymbol{\theta}})=\mathbf{h}(\mathbf{w}_i,\hat{\boldsymbol{\theta}})-N^{-1}\sum_i\mathbf{h}(\mathbf{w}_i,\hat{\boldsymbol{\theta}})$ 之上，并且对于建立在包括 OIR 检验基础的 $\hat{\boldsymbol{\theta}}$ 上的统计量来说，可获得渐近精炼。

霍罗维茨(Horowitz, 1988)对最小距离估计量(参见 6.7 节)做出了类似的重新中心化。然后，他把该自助法应用于 6.3.5 节中曾讨论过的奥尔顿吉和西格尔(Altonji and Segal, 1996)的协方差结构例子上。

一种可供选择的调整是由布朗和纽韦(Brown and Newey, 2002)提出的，它没有进行重新中心化，却以随不同观测值而变化的而非利用等于 $1/N$ 权数的概率再抽样观测 $\mathbf{w}_i$。特别地，设 $\Pr[\mathbf{w}^*=\mathbf{w}_i]=\hat{\pi}_i$，其中，$\hat{\pi}_i=(1+\hat{\boldsymbol{\lambda}}'\hat{\mathbf{h}}_i)$，$\hat{\mathbf{h}}_i=\mathbf{h}(\mathbf{w}_i,\hat{\boldsymbol{\theta}})$，而 $\hat{\boldsymbol{\lambda}}$ 使 $\sum_i\ln(1+\hat{\boldsymbol{\lambda}}'\hat{\mathbf{h}}_i)$ 最大化。其动机是，概率 $\hat{\pi}_i$ 等同于求解 $\sum_i\ln\pi_i$ 关于 $\pi_1,\cdots,\pi_N$ 的最小化(参见 6.8.2 节)的经验似然(EL)问题，使得约束 $\sum_i\pi_i\hat{\mathbf{h}}_i=\mathbf{0}$ 且 $\sum_i\pi_i=1$。因此，GMM 估计量的这种经验似然自助法(**empirical likelihood bootstrap**)利用了约束 $\sum_i\hat{\pi}_i\hat{\mathbf{h}}_i=\mathbf{0}$。

不过，一旦设 $\hat{\boldsymbol{\theta}}$ 表示 EL 估计量而不是 GMM 估计量，人们可以从开始就直接利用 EL 进行研究。布朗和纽韦(Brown and Newey, 2002)的方法优点是，它避开了对 EL 估计量进行计算的更具挑战性的问题。相反，人们只需要 GMM 估计量，并求解最小化 $\sum_i \ln(1+\hat{\boldsymbol{\lambda}}'\hat{\mathbf{h}}_i)$ 的凹规划问题。

11.6.5 *非参数回归*

非参数密度与回归估计量都以比 $\sqrt{N}$ 小的速率收敛，并且是渐近有偏的。这会使诸如置信区间推断错综复杂(参见 9.3.7 节与 9.5.4 节)。

我们考察 $m(x_0)=\mathrm{E}[y|x=x_0]$ 的核回归估计量 $\hat{m}(x_0)$，其中，观测值 (y,x) 是 iid 的，尽管允许出现条件异方差性。由霍罗维茨(Horowitz, 2001, 第 3 204 页)知，渐近中枢统计量是：

$$t=\frac{\hat{m}(x_0)-m(x_0)}{s_{\hat{m}(x_0)}}$$

其中，$\hat{m}(x_0)$ 表示具有带宽 $h=o(N^{-1/3})$，而不是最优 $h^*=O(N^{-1/5})$ 的光滑不足核回归估计量，同时：

$$s^2_{\hat{m}(x_0)}=\frac{1}{Nh[\hat{f}(x_0)]^2}\sum_{i=1}^{N}(y_i-\hat{m}(x_i))^2K\left(\frac{x_i-x_0}{h}\right)^2$$

其中，$\hat{f}(x_0)$ 表示密度 $f(x)$ 在 $x=x_0$ 处的核估计值。成对自助法再抽样 (y^*,x^*)，从而形成 $t_b^*=[\hat{m}_b^*(x_0)-m(x_0)]/s^*_{\hat{m}(x_0),b}$，其中，$s^*_{\hat{m}(x_0),b}$ 是利用自助法样本核估计值 $\hat{m}_b^*(x_i)$ 与 $\hat{f}_b^*(x_0)$ 而计算出来的。于是，11.2.7 节的分位数 t 置信区间提供了渐近精炼。对于对称置信区间或 α 水平上的对称检验来说，其误差是 $o(N^{-1}h)$，而不是利用一阶渐近近似的 $O(N^{-1}h)$。

有关这种自助法的几种变形是可行的。偏倚不是利用光滑不足，而是可直接通过估计 9.5.2 节曾给出的偏倚而得以剔除。同理，9.5.2 节给出的方差项不是利用 $s^2_{\hat{m}(x_0)}$，而是直接剔除。

亚特丘(Yatchew, 2003)曾经给出关于非参数回归与半参数回归中实施自助法的详细内容。

11.6.6 *非光滑估计量*

由 11.4.2 节知，自助法假定估计量与统计量具有光滑性。除此以外，自助法可能没有提供渐近精炼，而且甚至是无效的。

举例来说，我们考察 LAD 估计量及其对二值数据的推广。LAD 估计量(参见 4.6.2 节)具有目标函数 $\sum_i |y_i-\mathbf{x}_i'\boldsymbol{\beta}|$，该目标函数具有不连续的一阶导数。自助法可以提供有效的渐近近似，却不能提供渐近精炼。对二值结果来说，LAD 估计量可以推广到曼斯基(Manski, 1975)的最大得分估计量上(参见 14.7.2 节)。就此估计量而言，自助法甚至是不一致的。

在这些例子中，具有渐近精炼的自助法可通过利用估计量的原始目标函数的光滑形式来获得。例如，14.7.2 节将要阐述霍罗维茨(Horowitz, 1992)的光滑最

大得分估计量。

11.6.7 *时间序列*

自助法依赖于从 iid 分布中所进行的再抽样。因此,时间序列数据表现出明显的因相依性而引起的问题。

对于含有 ARMA 误差结构以及从基本的白噪声误差中进行再抽样的线性模型,自助法很容易实施。举一个例子,假定 $y_t=\beta x_t+u_t$,其中,$u_t=\rho u_{t-1}+\varepsilon_t$,而 ε_t 表示白噪声。然后,已知估计值 $\hat{\beta}$ 与 $\hat{\rho}$,我们就能递推地计算残差为 $\hat{\varepsilon}_t=\hat{u}_t-\hat{\rho}\hat{u}_{t-1}=y_t-x_t\hat{\beta}-\hat{\rho}(y_{t-1}-x_{t-1}\hat{\beta})$。一旦对这些残差进行自助法得出 $\hat{\varepsilon}_t^*$, $t=1,\cdots,T$,然后,递推计算 $\hat{u}_t^*=\rho\hat{u}_{t-1}^*+\hat{\varepsilon}_t^*$,从而 $y_t^*=\hat{\beta}x_t+\hat{u}_t^*$。于是,$y_t^*$ 对 x_t 进行回归,具有AR(1)误差。一个早期例子是由弗里德曼(Freedman, 1984)提出的,他对通过 2SLS 所估计的动态线性联立方程回归模型进行自助法。已知线性性,联立性会引发一点问题。模型的动态特性可借助于递推地构造 $\mathbf{y}_t^*=f(\mathbf{y}_{t-1}^*,\mathbf{x}_t,\mathbf{u}_t^*)$而得以处理,其中,$\mathbf{u}_t^*$ 表示通过从 2SLS 结构方程残差中进行再抽样而得到,并且 $\mathbf{y}_0^*=\mathbf{y}_0$。然后,对每个自助法样本实施 2SLS。

这种方法假定基本误差是 iid 的。对于不含 ARMA 设定的一般相关数据来说,例如,非平稳数据,可使用 11.5.2 节曾阐述的移动分块自助法。

为了检验单位根或协整,由于检验统计量的特性在单位根处会不连续地变化,所以应用自助法时需要特别小心谨慎。例如,参见李和马达拉(Li and Maddala, 1997)。尽管在这种情况下实施有效自助法是可能的,但迄今为止,这些自助法没有提供渐近精炼。

11.7 应用研究

应用研究者在借助于其他一些方法很难进行推断的情况下,不含渐近精炼的自助法就是一种相当有用的工具。这需要随着利用软件包和实践者工具箱而变。迄今为止,自助法的一种最普遍的应用是,对需要执行沃尔德假设检验的标准误差进行计算。一些例子包括异方差性稳健和面板稳健的推断、关于两步估计量的推断,以及对估计量变换的推断。其他一些潜在应用包括,对 m 检验统计量的计算,比如豪斯曼检验。

自助法能额外地提供渐近精炼。许多蒙特卡罗研究表明,在有限样本中相当标准的程序执行欠佳。存在着潜在的自助法精炼的应用,但目前还未实现。在一些情况下,这能改进现有的推断,比如在含有异方差的可加误差的模型中使用原始自助法。在另外一些情况下,它将激励对目前未充分利用方法的更多使用。特别地,具有良好小样本性质的设定检验能借助于对容易计算的辅助回归进行自助法而得以实施。

自助法的应用存在两个障碍。首先,自助法不总是装入统计软件包。这将会随着时间而变化,而且倘若软件包括循环且有能力保存回归输出,通过自助法建立代码就不再困难。其次,存在一些奥妙之处。渐近精炼需要使用渐近中枢统计量,

而且最简单的自助法假定 iid 数据，以及估计量与统计量的光滑性。这会涵盖一大类应用，但不包括所有应用。

11.8 文献注释

自助法是埃弗龙(Efron, 1979)针对 iid 数据提出的。辛格(Singh, 1981)、比克尔和弗里德曼(Bickel and Freedman, 1981)都曾提出早期的理论。一个好的入门统计学研究是由埃弗龙和蒂布沙兰尼(Efron and Tibsharani, 1993)给出的，而更高等研究则由霍尔(Hall, 1992)给出。对回归推广的情况很早就考虑过；例如，参见弗里德曼(Freedman, 1984)。最近十年来，经济计量学家进行大量研究工作。霍罗维茨(Horowitz, 2001)的综述是非常综合性的，而布朗斯通和卡齐米(Brounstone and Kazimi, 1998)综述则是一个良好的补充，他们考察了许多经济计量学应用，以及麦金农(MacKinnon, 2002)所撰写的论文。

习 题

11-1 考察模型 $y=\alpha+\beta x+\epsilon$，其中，$\alpha$、$\beta$ 以及 x 都表示纯量，且 $\epsilon\sim\mathcal{N}[0, \sigma^2]$。生成满足 $\alpha=2$，$\beta=1$ 以及 $\sigma^2=1$ 的容量 $N=20$ 的样本，且假定 $x\sim\mathcal{N}[2, 2]$。我们想要在水平 0.05 上利用 t 统计量 $t=(\hat{\beta}-1)/\mathrm{se}[\hat{\beta}]$ 检验 H_0: $\beta=1$ 与 H_a: $\beta\neq1$。使用 $B=499$ 次自助法复制。

(a) 给定斜率估计值 $\hat{\beta}$，通过 OLS 估计该模型。

(b) 运用成对自助法计算标准误差，并将之与最初样本估计值进行比较。使用自助法标准误差检验 H_0。

(c) 运用含有渐近精炼的成对自助法检验 H_0。

(d) 运用残差自助法计算标准误差，并将之与最初样本估计值进行比较。使用自助法标准误差检验 H_0。

(e) 使用含有渐近精炼的残差自助法检验 H_0。

11-2 依照下述 dgp 生成容量为 20 的样本。借助于 $x_1\sim\chi^2(4)-4$ 与 $x_2\sim 3.5+\mathcal{U}[1, 2]$来生成两个回归元；误差是出自以概率 0.3 满足正态分布 $u\sim\mathcal{N}[0, 25]$且以概率 0.7 满足 $u\sim\mathcal{N}[0, 5]$的混合分布；而因变量 $y=1.3x_1+0.7x_2+0.5u$。

(a) 通过 OLS 估计模型。

(b) 假定我们对数据估计 $\gamma=\beta_1+\beta_2^2$ 感兴趣。利用最小二乘法来对这个量进行估计。使用 δ 方法获得该函数的近似标准误差。

(c) 随后，利用成对自助法对 $\hat{\gamma}$ 的标准误差进行估计。把这个值与来源于(b)的 $\mathrm{se}[\hat{\gamma}]$加以比较，并解释其差异。对于自助法来说，使用 $B=25$ 且 $B=200$。

(d) 现在，利用 $B=999$ 的成对自助法在 0.05 水平上检验 H_0: $\gamma=1.0$。实施含有渐近精炼的自助法与没有渐近精炼的自助法。

11-3 对健康消费支出(y)的自然对数与总消费支出(x)的自然对数，使用源

于 4.6.4 节的 200 个观测值。求模型 $y=\alpha+\beta x+u$ 的 OLS 估计。使用 $B=999$ 的成对自助法。

(a) 求 $\hat{\beta}$ 的标准误差的自助法估计。

(b) 利用这个标准误差估计值来检验 $H_0: \beta=1$ 与 $H_a: \beta\neq1$。

(c) 在 u 是同方差的假设下，实施含有精炼的 $H_0: \beta=1$ 与 $H_a: \beta\neq1$ 的自助法检验。

(d) 如果 u 是异方差的，你在(c)中所用的方法会怎样呢？检验还是渐近有效的吗？检验提供渐近精炼时会是这样吗？

(e) 运用自助法求 β 的偏倚修正估计值。

12 基于模拟的方法

12.1 引　论

前几章曾经阐述，非线性方法并不要求估计量有闭形式解。不过，非线性方法却紧密地依赖于解析处理性。尤其是，假定估计量的目标函数具有闭形式表达式，同时估计量的渐近分布建立在估计方程的线性化基础上。

在本章，我们将阐述基于模拟的估计方法。第 5 章对 ML 估计的研究已假定，密度 $f(y|\mathbf{x},\boldsymbol{\theta})$ 具有闭形式表达式。如果不存在闭形式解，而当我们使用 $f(y|\mathbf{x},\boldsymbol{\theta})$ 的一个良好近似 $\hat{f}(y|\mathbf{x},\boldsymbol{\theta})$ 去建立似然函数时，极大似然估计或许还是可行的。缺乏密闭形式表达式的普遍原因是，$f(y|\mathbf{x},\boldsymbol{\theta})$ 定义中存在不易处理的期望。例如，在随机系数模型中，对随机参数进行积分并将之去掉很困难。倘若期望用蒙特卡罗近似来代替，则得到的估计量被称为基于模拟的估计量。类似的模拟方法能应用于建立在矩基础上的矩估计方法，诸如条件均值，原因在于没有闭形式解。在矩方法情况下，用模拟方法获得一致参数估计是可行的，此时，比极大似然估计情况下为一致性而必需的模拟要更少一些。

这些估计方法都是密集计算的，因为它们大量运用蒙特卡罗抽样方法。运用蒙特卡罗方法，将引起近似的准确性、计算有效性以及使用这类近似估计量的抽样性质问题。

12.2 节给出基于模拟估计的动机例子。12.3 节涵盖计算积分的基础，其中会提及连续随机变量是一个积分的问题。12.4 节与 12.5 节阐述极大模拟似然估计与模拟的基于矩估计；12.6 节研究间接推断。这些估计量需要模拟器[1]（simulators），详细内容则在 12.7 节阐明，而伪随机数在 12.8 节加以详细阐述。

12.2 例　子

我们考察下述例子，给定回归元 $\mathbf{x}$ 与参数 $\boldsymbol{\theta}$ 时，y 的条件密度是一个积分：

〔1〕 又称为模拟式或模拟装置。——译者注

$$f(y|\mathbf{x},\boldsymbol{\theta})=\int h(y|\mathbf{x},\boldsymbol{\theta},\mathbf{u})g(\mathbf{u})d\mathbf{u} \tag{12.1}$$

其中，$h(\cdot)$与$g(\cdot)$的函数形式均已知，而 $\mathbf{u}$ 表示随机变量，不一定为误差项，它需要通过积分而去掉。如果此积分不存在解析解，从而似然函数没有闭形式表达式，那么这就成为使用基于模拟的估计方法的根据。

12.2.1 *随机参数模型*

随机参数模型(**random parameter model**)或随机系数模型(**random coefficients model**)允许回归系数依据某个分布随不同个体而变化。一种完全参数随机参数模型，设定以回归元 $\mathbf{x}_i$ 与给定参数 γ_i 为条件的因变量 y_i 具有条件密度 $f(y_i|\mathbf{x}_i,\gamma_i)$，其中，$\gamma_i$ 是 iid 的，其密度为 $g(\gamma_i|\boldsymbol{\theta})$。推断建立在以 $\mathbf{x}_i$ 与给定 $\boldsymbol{\theta}$ 为条件的 y_i 的密度基础上，即：

$$f(y|\mathbf{x},\boldsymbol{\theta})=\int f(y|\mathbf{x},\gamma)g(\gamma|\boldsymbol{\theta})d\gamma \tag{12.2}$$

除了在一些特殊情况下，此积分将没有闭形式解。一种普遍设定是，假定正态分布随机参数，满足 $\gamma_i\sim\mathcal{N}[\boldsymbol{\mu},\boldsymbol{\Sigma}]$。于是，$\gamma_i=\boldsymbol{\mu}+\boldsymbol{\Sigma}^{-1/2}\mathbf{u}_i$，其中，$\mathbf{u}_i\sim\mathcal{N}[\mathbf{0},\mathbf{I}]$，并能用式(12.1)的形式重新写出式(12.2)，其中，$\boldsymbol{\theta}$ 表示包含 $\boldsymbol{\mu}$ 与 $\boldsymbol{\Sigma}$ 的独特分量的向量，而 $g(\mathbf{u})$表示 $\mathcal{N}[\mathbf{0},\mathbf{I}]$的密度。

随机参数模型的一个简单例子是被忽略异质性(**neglected heterogeneity**)。于是，经常恰好有一个参数即通常的截距被假定成随机的，因而积分是一维的，这很容易在数值上加以近似。不过，更一般地讲，积分维数可能是高维的。

随机参数与不可观测异质性的一些重要例子包括：(1)多项式 logit 模型中的服从正态分布随机参数(随机参数 logit 模型，参见第 15 章)；(2)威布尔持续期限模型中的服从伽玛分布不可观测异质性(参见第 19 章)；(3)泊松计数模型中的服从伽玛分布不可观测异质性(参见第 20 章)；以及(4)面板数据模型中的特定个体随机效应(参见第 21 章)。在对异质性分布进行积分之后，对于例子 3 与例子 4 所得到的边缘密度的闭形式解在正态性下，对线性模型来说都是可利用的。可是，对于例子 1、例子 2 以及例子 4 的许多非线性应用来说，没有闭形式解可利用。

12.2.2 *受限因变量模型*

受限因变量(**limited dependent variable**，记为 LDV)是指，因变量由于删失或截取而仅仅在其一部分取值范围内才是可观测的。于是，可观测变量的密度会涉及不可能具有闭形式表达式的积分。

受限因变量的一类重要例子是离散选取模型(**discrete choice models**)，第 14 章与第 15 章将详述此类模型。我们这里介绍离散选择模型，它们是基于模型估计的经济计量学文献所关注的内容。

举一个例子，考察消费者在三种互不相交的商品之间选择其一，诸如三种各不相同的耐用商品，消费者从中选择唯一商品。假定消费者对效用求最大化，并设可供选取的商品 1、商品 2、商品 3 的效用分别用 U_1、U_2、U_3 给出。效用 U_1、U_2 以及

U_3 是不可观测的。然而,我们仅仅可以观察到依赖于被选取的商品离散结果变量 $y=1$、2 或者 3。

假定可选商品 1 被选上,因为它具有最高的效用。于是,其概率质量函数是 $p_1=\Pr[y=1]$,其中:

$$\begin{aligned}p_1&=\Pr[U_1-U_2\geqslant 0,\ U_1-U_3\geqslant 0]\\&=\Pr[(\mathbf{x}_1-\mathbf{x}_2)'\boldsymbol{\beta}+\varepsilon_1-\varepsilon_2\geqslant 0,\ (\mathbf{x}_1-\mathbf{x}_3)'\boldsymbol{\beta}+\varepsilon_1-\varepsilon_3\geqslant 0]\end{aligned}$$

如果我们做出共同假设(参见 15.5.1 节),即 $U_j=\mathbf{x}_j'\boldsymbol{\beta}+\varepsilon_j$,$j=1,2,3$,其中,回归元 $\mathbf{x}$ 测量了这三种商品的不同属性,而误差 ε 可在 $(-\infty,\ \infty)$ 上变化。一旦定义 $u_1=U_1-U_2$ 且 $u_2=U_1-U_3$,有:

$$p_1=\int_0^{\infty}\int_0^{\infty}g(u_1,u_2)du_1du_2 \tag{12.3}$$

其中,$g(u_1,u_2)$,或更正式地,$g(u_1,u_2|\mathbf{x},\boldsymbol{\theta})$ 表示 (u_1,u_2) 的二变量密度,或者等价地:

$$p_1=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}1[u_1\geqslant 0,\ u_2\geqslant 0]g(u_1,u_2)du_1du_2 \tag{12.4}$$

其中,$1[A]$ 表示指示变量函数,当事件 A 发生,$1[A]$ 就等于 1,否则等于 0。

积分式(12.4) 具有式(12.1) 的形式。由于积分仅对 (u_1,u_2) 的一部分范围进行[参见式(12.3)],所以不可能存在闭形式解,即使我们知道,如果积分在 (u_1,u_2) 整个范围进行,那么 $\iint g(u_1,u_2)du_1du_2=1$。

特别地,当误差 ε 服从正态分布,如同多项式 probit 模型(**multinominal probit model**),积分式(12.3)是在二变量正态分布的正象限进行。p 不存在闭形式解,因而对于密度 $f(y|\mathbf{x},\boldsymbol{\theta})$ 来说,不存在容易处理的表达式。在实际应用中,积分维数可能非常高,用数值形式加以近似很难,因为对于在 m 个互斥可供选择的情况之间选取来说,积分具有 $m-1$ 维数。一直到发展出基于模拟的估计量,研究者才使用含有 $m\leqslant 4$ 的模型,或者选择其他的误差分布,比如导致更强约束的多项式 logit 模型。

12.2.3 *ML 估计*

为了简单起见,考察 MLE。假定不同的观测值具有独立性,同时 y 具有条件密度 $f(y|\mathbf{x},\boldsymbol{\theta})$。

前面两个例子中的新困难是,ML 估计行不通,因为 $f(y|\mathbf{x},\boldsymbol{\theta})$ 不存在闭形式表达式,是由不能加以简化的积分来定义的。相反,我们用数值形式近似 $\hat{f}(y|\mathbf{x},\boldsymbol{\theta})$ 来代替此积分,然后对:

$$\ln\hat{L}_N(\boldsymbol{\theta})=\sum_{i=1}^{N}\ln\hat{f}(y_i|\mathbf{x}_i,\boldsymbol{\theta})$$

求关于 $\boldsymbol{\theta}$ 的极大值。此估计量将是一致的,并且如果 $\hat{f}(y|\mathbf{x},\boldsymbol{\theta})$ 是 $f(y|\mathbf{x},\boldsymbol{\theta})$ 的一个良好近似,它就具有与 MLE 相同的渐近分布。

所得到的一阶条件通常是非线性的，并通过迭代法来求解。因为 $\hat{f}(y_i|\mathbf{x}_i,\boldsymbol{\theta})$ 随 i 与 $\boldsymbol{\theta}$ 而变化，故利用数值导数进行梯度计算将需要至少计算 Nqr 次，其中，N 表示样本量，q 表示 $\boldsymbol{\theta}$ 的维数，而 r 表示迭代次数。例如，对于具有 1 000 个观测值、10 个参数以及 50 次迭代来说，至少计算 500 000 次函数。

非线性模型所需要的这种标准计算，现在要用为计算对积分 $f(y|\mathbf{x},\boldsymbol{\theta})$ 适宜的近似而需的计算次数去乘。很明显，人们希望获得计算次数相对少一些的方法。

12.2.4 贝叶斯方法

第 13 章将给出对贝叶斯方法的单独研究。这些方法包括类似于式(12.2)的积分计算，但要进一步加以计算，从而得到参数的(后验)分布，而不是诸如极大似然估计的点估计。

12.3 积分计算基础

我们考察积分(**integral**)：

$$I=\int_a^b f(x)dx \tag{12.5}$$

其中，$f(\cdot)$在$[a, b]$上是连续的，而且积分的界限不需要是有限的，因此，$a=-\infty$，并且或者 $b=\infty$ 是可能的。在本节，x 最初表示纯量，并表示可用积分去掉的变量。在回归应用中，积分经常是针对向量的，该向量记为 $\mathbf{u}$，进而 $\mathbf{x}$ 表示回归元[参见式(12.1)]。假定积分存在，即使积分发散，需要一种重要的限定条件得到 I 的有限估计值，该限定条件用于核对其近似方法。

首先，我们阐述对于低维数积分有用的数值积分或求积分。这通过蒙特卡罗积分来完成，对于高维数积分来说会更好地起作用，这也是本章关注的内容。

本节内容与实施基于模拟的估计有关；因此，一些读者可能愿意在讨论 12.4 节至 12.6 节之后阅读它。

12.3.1 确定性数值积分

积分能被解释成为对面积或者体积的测量。确定性数值积分或求积分(**deterministic numerical integration or quadrature**)是用一系列较小体积的切片加起来代替体积。正式地讲，这会涉及对被积函数在几个点上进行计算，同时对这些值取加权和。确定前缀用于表明，对积分近似的这一方法不需要进行模拟。

辛普森法则

由积分定义：

$$I=\lim_{\Delta x_j\to 0}\sum_{j=1}^{N}f(x_j)\Delta x_j \tag{12.6}$$

其中，x 的范围$[a, b]$被分割成$(n+1)$个点，$x_0<x_1<\cdots<x_n$，并且 $n\to\infty$。一些标准近似方法是，对有限 n 提供更准确的式(12.6)的精炼式。我们对等距点阐述其结果，尽管这些结果能被推广到对那些不等距点的计算上。为了简单起见，假定

$f(x)$在极限点 a 与 b 处可以计算。

中点法则(**midpoint rule**)是指在区间$[x_{j-1},x_j]$的中点 $\bar{x}_j=\frac{1}{2}(x_{j-1}+x_j)$进行计算,然后对底为$(b-a)/n$ 而高为 $f(\bar{x}_j)$的 n 个矩形进行求和。因而,I 通过:

$$\hat{I}_M=\sum_{j=1}^{n}\frac{b-a}{n}f(\bar{x}_j) \tag{12.7}$$

来逼近。梯形法则(**trapezoidal rule**)是对 $f(x_{j-1})$与 $f(x_j)$之间的连接直线加以改进,然后对底为$(b-a)/n$ 而平均高为$(f(x_{j-1})+f(x_j))/2$ 的 n 个梯形进行求和。因而,I 通过:

$$\hat{I}_T=\sum_{j=1}^{n}\frac{b-a}{n}\frac{f(x_{j-1})+f(x_j)}{2} \tag{12.8}$$

来逼近。辛普森法则(**Simpson's rule**)在三个相继连接点 $f(x_{j-1})$、$f(x_j)$以及 $f(x_{j+1})$之间使用二次曲线,而梯形法则在相继连接点之间使用直线。从而得到下述近似:

$$\hat{I}_S=\sum_{j=0}^{n}\frac{(b-a)}{3n}w_j f(x_j) \tag{12.9}$$

其中,n 表示偶数,除 $w_0=w_n=1$ 之外,当 j 为奇数时,$w_j=4$,而当 j 为偶数时,$w_j=2$。

这些近似误差界限会作为积分范围 $b-a$ 的幂函数而增大,并作为积分次数的幂函数而减小。对于辛普森法则来说,$|I_S-I|\leqslant M_4(b-a)^5/180n^4$,其中,$M_4$ 表示 x 在$[a,b]$上四阶导数的最大绝对值。对于梯形法则来说,$|I_T-I|\leqslant M_2(b-a)^3/12n^2$,其中,$M_2$ 表示 x 在$[a,b]$上二阶导数的最大绝对值。很明显,积分次数需要随着 x 范围而增大,而且人们应检验积分次数的敏感性。

辛普森法则及其有关的法则,对于有限区间上的定积分起着良好的作用。可是,很明显,对于不定积分来说却产生了问题,因为出现要计算尾部的问题,例如,假定$[a,b]=[0,\infty)$。于是,在选取 x_n 时存在一种权衡,因为上界 x_n 应是很大的,从而计算点之间的距离也很大。至少人们应去检验对 x_n 增大的敏感性。

高斯求积法

高斯求积法(**Gaussian quadrature**)是高斯在 1814 年提出的,它是一种可供选择的以数值积分命名的积分。它提供了对计算点 x_j 不再等距的一种良好的选取法则,同时尤其有助于计算不定积分。

首先,把式(12.5)重新写成:

$$I=\int_c^d w(x)r(x)dx \tag{12.10}$$

其中,$w(x)$通常依赖于 x 的范围,并是下述三种函数之一:高斯—埃尔米特(Gauss-Hermite)积分,设 $w(x)=e^{-x^2}$,并用于$[c,d]=(-\infty,\infty)$;高斯—拉盖尔(Gauss-Laguerre)积分,设 $w(x)=e^{-x}$,并用于$[c,d]=(0,\infty)$;高斯—勒让德(Gauss-

Legendre)积分，设 $w(x)=1$，并用于$[c,d]=[-1,1]$。

在最简单的情况下，式(12.10)通过定义 $r(x)=f(x)/w(x)$，从式(12.5)中获得。更一般地讲，可能要求 x 的变换，例如式(12.5)的范围$[2,\infty)$变成式(12.10)的$[0,\infty)$。一些方法，允许使用者直接提供 $f(x)$与积分的范围，并自动处理任何必需的变换。

高斯求积法是通过加权和：

$$\hat{I}_G = \sum_{j=1}^{m} w_j r(x_j) \tag{12.11}$$

逼近积分式(12.10)，其中，m 由研究者选取；m 个计算点 x_j 与加权 w_j 在诸如阿布拉莫维茨和斯特古(Abramowitz and Stegun，1971)的书中或由普雷斯等人(Press et al.，1993)提供的计算机程序中都可找到。

支撑近似的理论是建立在 $w(x)$正交多项式(**orthogonal polynomial**)的基础上，记为 $p_j(x)$，$j=0,\cdots,m$，它满足：

$$\int_c^d w(x)p_j(x)p_k(x)dx = 0, \quad j \neq k, \quad j,k = 0,\cdots,m$$

另外，当$\int_c^d w(x)p_j^2(x)dx = 1$时，就称该多项式是正交的。如果 $r(x)$ 是阶数为 $2m-1$ 或小于 $2m-1$ 的多项式，那么近似式(12.11) 是准确的，因此若式(12.10) 的 $r(x)$ 是由阶数 $2m-1$ 的多项式来很好地逼近，则近似就行得通。对计算点 m 个数的一种良好选取需要通过试验来定，但许多应用只使用 20 或 30。

举一个例子，考察高斯—埃尔米特求积法(**Gauss - Hermite quadrature**)，由于积分经常是在$(-\infty, \infty)$上进行，故该方法在经济计量学里普遍使用。对于 $w(x)=e^{-x^2}$ 来说，正交多项式 $p_j(x)$是埃尔米特多项式 $H_j(x)$，其中，正交形式是利用递归式 $H_{j+1}(x) = \sqrt{2/(j+1)}\,xH_j(x) - \sqrt{j/(j+1)}\,H_{j-1}(x)$ 生成的，$j=1,\cdots,m$，$H_{-1}=0$，并且 $H_0=\pi^{-1/4}$。所得出的 m 个横坐标 x_j 作为 $H_m(x)=0$ 的 m 个根，同时对于正交埃尔米特多项式来说，权数 $w_j=1/[jH_{j-1}(x_j)^2]$。正如已注意到的，给定 m，x_j 与 w_j 在表格或计算机编码中都是可以利用的。

对于定积分来说，高斯—勒让德求积法通常比辛普森法则实施得更好。不过，高斯求积法的实际优点是针对不定积分的。注意到，若积分在$(-\infty,\infty)$上计算，通过变量变换变成$(0,\infty)$上的积分是可行的，然后用高斯—拉盖尔求积法而不是高斯—埃尔米特求积法。

存在另外一些计算积分的确定性方法，包括拉普拉斯近似法[蒂尔尼、卡斯和卡登那(Tierney，Kass，and Kadane，1989)]。

12.3.2 *通过直接蒙特卡罗抽样积分*

蒙特卡罗积分为确定性数值积分提供了一种可供选择的方法。通常，对 $I = \int_a^b f(x)dx$ 的蒙特卡罗积分估计是：

$$\hat{I}_{\mathrm{MC}} = \sum_{s=1}^{S} f(x^s) \tag{12.12}$$

其中，$x^1,\cdots,x^s$ 表示 S 从范围$[a, b]$中均匀采样。与中点法则相比，我们在 S 个随机选取的点上而不是 n 个确定的中点上计算 $f(x)$。

我们关注一些回归的应用，诸如 12.2 节给出的那些例子。于是，由于想要获得期望值 $\mathrm{E}[h(x)]$，比如说，期望是关于随机变量 x 的，不妨设随机变量 x 具有 pdf $g(x)$，所以就产生了积分。在连续情况下，我们想要计算：

$$\mathrm{E}[h(x)] = \int_a^b h(x)g(x)dx \tag{12.13}$$

本章自始至终地假定，$\mathrm{E}[h(x)]<\infty$，也就是积分收敛。然后，$\mathrm{E}[h(x)]$可通过直接蒙特卡罗积分法（**direct Monte Carlo integral estimate**）得出估计：

$$\hat{I}_{\mathrm{DMC}} = \hat{\mathrm{E}}[h(x)] = S^{-1}\sum_{s=1}^{S} h(x^s) \tag{12.14}$$

其中，$\{x^s, s=1,\cdots,S\}$表示来自密度 $g(x)$的 S 个伪随机数的蒙特卡罗样本，这可利用稍后 12.8 节将给出的方法来得到。估计式(12.14)利用了来自密度 $g(x)$的 x 采样对 $h(x)$进行估计，而估计式(12.12)则利用如同式(12.12)中 x 的均匀抽取采样对 $h(x)g(x)$进行估计。式(12.14)的优点是，它能应用于不定积分，而且当 a 或 b 无界时，要获得式(12.12)的均匀采样就出现问题了。

估计值 $\hat{\mathrm{E}}[h(x)]$是函数 $f(\cdot)$在每一个随机采样 x^s 上计算值的平均。等价地，$\hat{\mathrm{E}}[h(x)]$就是随机变量 $h(x_s)$的平均，而且当 $S\to\infty$时，如果应用大数定律和中心极限定理，就可得出它的性质。此处，x^s 是 iid 的，因而 $h(x^s)$是 iid 的，由于 $\mathrm{E}[h(x)]$存在是已经假定的，所以我们可应用柯尔莫哥洛夫 LLN(参见附录 A，定理 A.8)。由此可得：

$$\text{当 } S\to\infty\text{时，}\ \hat{\mathrm{E}}[h(x)] \xrightarrow{p} \mathrm{E}[h(x)]$$

同理，由于 $h(x^s)$是 iid 的，一旦假定 $\mathrm{V}[h(x)]$存在，则 $\hat{\mathrm{E}}[h(x)]$的方差等于 $S^{-1}\mathrm{V}[h(x)]$。当 $S^{-1}\mathrm{V}[h(x^s)]$ 很小，这种近似对于适度的 S 大小来说可以是良好的。

12.3.3 *积分计算例子*

假定 $x\sim\mathcal{N}[0, 1]$，我们想要计算均值：

$$\mathrm{E}[x] = (\sqrt{2\pi})^{-1}\int_{-\infty}^{\infty} x\exp(-x^2/2)dx$$

以及矩 $\mathrm{E}[\exp(-\exp(x))]$，而该值被定义为下述积分：

$$\mathrm{E}[\exp(-\exp(x))] = (\sqrt{2\pi})^{-1}\int_{-\infty}^{\infty}\exp(-\exp(x))\ \exp(-x^2/2))dx$$

$\mathrm{E}[x]$ 的解析表达式存在，而且得到 $\mathrm{E}[x]=0$。与之相比，$\mathrm{E}[\exp(-\exp(x))]$ 的解析表达式却不存在。在寻找数值近似之前，首先证实此积分确实是收敛的。

由于 $\exp(-\exp(x))$ 是严格正的，且随着最大值 1 而单调递减，由此可得，$|\exp(-\exp(x))|<1$，因此 $E[\exp(-\exp(x))]=E[1]=1$，从而积分收敛。

这些一维积分很容易利用确定性数值近似来计算。例如，考察对 $x_0=-5$ 与 $x_{20}=5$ 之间具有 $n=20$ 的等距计算值的中点法则。于是，有：

$$\hat{E}[x] = (\sqrt{2\pi})^{-1}\sum_{j=1}^{20}\frac{10}{20}\bar{x}_i\exp(-\bar{x}_j^2/2)$$

$$\hat{E}[\exp(-\exp(x))] = (\sqrt{2\pi})^{-1}\sum_{j=1}^{20}\frac{10}{20}\exp(-\exp(\bar{x}_j))\exp(-\bar{x}_j^2/2)$$

其中，$\bar{x}_j=-5.25+j/2$。正如人们所料，当小数位数很多时，$\hat{E}[\exp(-\exp(x))]=0.381\,756\,56$。相反，当我们在 -10 与 10 之间令 $n=200$，后者估计值变化很小，一直到第 8 位小数。很明显，此处的确定性数值方法表现良好。

这些积分还可利用蒙特卡罗近似来计算，并满足：

$$\hat{E}[x] = \frac{1}{S}\sum_{s=1}^{S}x^s$$

$$E[\exp(-\exp(x))] = \frac{1}{S}\sum_{s=1}^{S}\exp(-\exp(x^s))$$

其中，x^s 表示从 $\mathcal{N}[0,1]$ 分布中得到 S 个抽取的第 s 个采样，而实施这类采样方法已在附录 B 中给出。表 12.1 对于模拟 S 的各种不同次数给出 $\hat{E}[x]$ 与 $\hat{E}[\exp(-\exp(x))]$ 的估计。注意到，当 $S\rightarrow\infty$ 时，此估计量趋于稳定，且分别趋向于它们的各自真实值 0 与 0.381 756 56，其中后者可通过确定性数值近似来获得。不过。当 $S=10^6$，估计值 $\hat{E}[x]$ 还在第 4 个小数位上不同于 0。这里，由于 $V[x^s]=1$，$V[\hat{E}[x]]=S^{-1}V[x^s]=1/S$，因此，甚至当 $S=10^6$，$\hat{E}[x]$ 的标准差是相对大的，即 0.001。具有较小方差的蒙特卡罗近似的一种可供选择方法将在 12.7 节给出。

表 12.1　蒙特卡罗积分：关于 x 标准正态的例子

S=模拟次数	$\hat{E}[x]$	$E[\exp(-\exp(x))]$
10	0.145	0.336
25	−0.209	0.435
50	0.050	0.369
100	−0.120	0.409
500	−0.059	0.398
1 000	0.005	0.382
10 000	−0.007	0.383
100 000	−0.000	0.382
1 000 000	−0.000	0.381

12.3.4 *较高维数积分*

较高维数的积分，可利用确定性积分或蒙特卡罗积分法来进行计算，当维数增大时，后一种方法更受人们喜欢。

确定性积分法可利用多元高斯求积分做得更好，或者如果积分极限不太复杂时，通过把 m 维积分简化成 m 个一维积分来计算，比如说，利用高斯求积法。不过，由式(12.6)积分定义知，很明显计算次数将以 m 幂增长。例如，对于一维积分来说，需要计算 20 次函数，而五维积分可能需要 5^{20} 次或 95 万亿次计算。当对每一个观测值都要计算，然后求和进行估计时，并不需要这样高的精度，但是其计算次数实际上随积分维数而增加。

可直接实施较高维数的蒙特卡罗积分法，只是把式(12.13)与式(12.14)中的 x 定义为一个向量，并从多元密度 $g(\mathbf{x})$ 中采样。很明显，不存在维数祸根。不过，人们应记住，如果被积函数具有强烈的峰值，那么简单蒙特卡罗积分法将不会起作用，而且一种可能情况是，这种峰值在较高维数中可能变得特别显著。特别地，对于 12.2.2 节的离散选择例子来说，式(12.4)的被积函数仅仅在(u, v)范围的一小部分上可能为非零，12.7 节将继续探讨其含义。此外，与从一元分布中采样相比，从多元分布中采样更加困难。

12.4 极大似然模拟估计

现在，考察当密度没有解析表达式可利用时，把这些想法应用于 ML 估计上。一个重要结果是，模拟能产生具有与 MLE 相同分布的估计量，倘若对每一个观测值来说，为计算密度而采样的模拟次数趋向于无穷大。

12.4.1 *模拟器*

假定观测值的条件密度 $f(y|\mathbf{x},\boldsymbol{\theta})$ 涉及不容易处理的积分。具体地讲，假定如同式(12.1)一样：

$$f(y_i|\mathbf{x}_i,\boldsymbol{\theta}) = \int h(y_i|\mathbf{x}_i,\boldsymbol{\theta},\mathbf{u}_i)g(\mathbf{u}_i)d\mathbf{u}_i \tag{12.15}$$

若不存在闭形式解，就需要对它进行估计。

$f(y_i|\mathbf{x}_i,\boldsymbol{\theta})$的直接模拟器[1]（**direct simulator**）是一种明显的蒙特卡罗积分法估计：

$$\hat{f}(y_i|\mathbf{x}_i,\mathbf{u}_{iS},\boldsymbol{\theta}) = \frac{1}{S}\sum_{s=1}^{S}h(y_i|\mathbf{x}_i,\boldsymbol{\theta},\mathbf{u}_i^s) \tag{12.16}$$

其中，$\mathbf{u}_{iS}$表示 S 个采样的 $\mathbf{u}_i^s$ 向量，$s=1,\cdots,S$，它们是从 $g(\mathbf{u}_i)$中独立抽样。这直接对 S 个采样求 $h(y_i|\mathbf{x}_i,\boldsymbol{\theta},\mathbf{u}_i^s)$的平均数。由 12.3.2 节知，当采样次数 $S\to\infty$时，$\hat{f}_i$ 关于 f_i 是无偏的且关于 f_i 是一致的。

除直接模拟器以外，可使用其他一些模拟器，这些将在 12.7 节加以详述。例如，倘若采样还具有边缘分布 $g(\mathbf{u}_i)$，就允许采样之间存在相关，这些将会产生估计值 $\hat{f}_i$，对于有限采样次数来说，$\hat{f}_i$ 较好地逼近 f_i。于是，更一般地讲，$f(y_i|\mathbf{x}_i,\boldsymbol{\theta})$ 的蒙特卡罗估计是：

---

[1] 又称为直接模拟式。——译者注

$$\hat{f}(y_i \mid \mathbf{x}_i, \mathbf{u}_{iS}, \boldsymbol{\theta}) = \frac{1}{S}\sum_{s=1}^{S}\tilde{f}(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}, \mathbf{u}_i^s) \tag{12.17}$$

其中，$\mathbf{u}_i^s$ 表示 S 个采样，其边缘密度为 $g(\mathbf{u}_i)$，$s=1,\cdots,S$，但对于不同的 s 来说，不一定是独立的。为了运用该模拟器，当 $S\to\infty$时，$\hat{f}_i \xrightarrow{p} f_i$。子模拟器(**subsimulator**) $\tilde{f}(\cdot)$被称为是无偏模拟器(**unbiased simulator**)，如果它满足下述性质：

$$\mathrm{E}[\tilde{f}(y \mid \mathbf{x}, \boldsymbol{\theta}, \mathbf{u}^s)] = f(y \mid \mathbf{x}, \boldsymbol{\theta}) \tag{12.18}$$

模拟器的一个值得拥有的性质是，$\hat{f}_i$ 在 $\boldsymbol{\theta}$ 上是可微的，因此，标准迭代梯度法能用于计算 $\boldsymbol{\theta}$ 的估计。为了剔除因模拟引发的“振动”(**chatter**)，并确保数值收敛，用于构造 $\hat{f}_i$ 的基本蒙特卡罗抽样不应该重复抽样，因为 $\boldsymbol{\theta}$ 对于不同迭代会变化。

12.4.2 *MSL 估计量*

已知对不同 i 具有独立性，极大似然估计量 $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$是对 $\ln L_N(\boldsymbol{\theta}) = \sum_{i=1}^{N}\ln f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta})$求极大值。可是，模拟极大似然(**maximum simulated likelihood**, MSL)估计量 $\hat{\boldsymbol{\theta}}_{\mathrm{MSL}}$是对建立在密度模拟估计基础上的对数似然求极大值，或者：

$$\ln \hat{\mathrm{L}}_N(\boldsymbol{\theta}) = \sum_{i=1}^{N}\ln \hat{f}(y_i \mid \mathbf{x}_i, \mathbf{u}_{iS}, \boldsymbol{\theta}) \tag{12.19}$$

其中，模拟器 $\hat{f}(\cdot)$已在式(12.17)中定义。如果 $\hat{f}(\cdot)$在 θ 上是可微的，那么利用第 10 章的标准梯度法，运用解析导数或数值导数来计算 $\hat{\boldsymbol{\theta}}_{\mathrm{MSL}}$。

12.4.3 *MSL 估计量的分布*

由 5.3.2 节概述的一般一致性证明法知，当逼近目标函数 $N^{-1}\ln \hat{\mathrm{L}}_N(\boldsymbol{\theta})$具有与最初目标函数 $N^{-1}\ln\hat{L}_N(\boldsymbol{\theta})$相同的概率极限，MSL 估计量将具有与 ML 估计量同样的概率极限。如果 $\ln \hat{f}_i - \ln f_i \xrightarrow{p} 0$，那么就是这种情况，同样地，若 $\hat{f}_i - f_i \xrightarrow{p} 0$，当 $S\to\infty$时，就是这种情况。

即使 MSL 估计量是一致的，可能出现：模拟误差使 MSL 估计量的方差比 ML 估计量的要更大一些。举一个例子，在我们给出下述命题，即 MSL 估计量为完全有效的那种条件下对条件的正式叙述，而此命题是对古里耶克斯和蒙福特定理(Gouriéroux and Monfort, 1991)重新表述而形成的。

命题 12.1(MSL 估计量的分布)[古里耶克斯和蒙福特(Gouriéroux and Monfort,1991)] 假定下述条件：

(i) 数据来自具有条件密度 $f(y \mid \mathbf{x}, \boldsymbol{\theta}_0)$的 dgp 的简单随机样本，并满足正则条件，因此，ML 估计量是一致的且渐近正态的，其极限方差矩阵为$\mathbf{A}^{-1}(\boldsymbol{\theta}_0)$，其中：

$$\mathbf{A}(\boldsymbol{\theta}_0) = -\operatorname{plim}\left[N^{-1}\sum_{i=1}^{N}\frac{\partial^2 \ln f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}_0}\right]$$

(ii) 密度 f 可利用式(12.17)的模拟器来估计，并且 $\hat{f}$ 关于 f 是无偏的。

于是，当 S, $N\to\infty$ 且 $\sqrt{N}/S\to 0$，由式(12.19)定义的极大似然模拟(**maximum**

simulated likelihood)估计量是渐近等价于 ML 估计量，并且它服从极限正态分布，满足：

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{\mathrm{MSL}}-\boldsymbol{\theta}_0)\xrightarrow{d}\mathcal{N}[0,\ \mathbf{A}^{-1}(\boldsymbol{\theta}_0)] \qquad (12.20)$$

在比较弱的条件下，即 $S,N\rightarrow\infty$时，MSL 估计量实际上是一致的。例如，如果对某个常数 a 而言，$S=N^{0.4}/a$，那么这就得到满足。然而，$\sqrt{N}/S=aN^{0.1}\rightarrow\infty$，因而按照命题 12.1，MSL 估计量不是完全有效的。由通常一阶泰勒级数展开式，$\sqrt{N}(\hat{\boldsymbol{\theta}}_{\mathrm{MSL}}-\boldsymbol{\theta}_0)$的极限分布是 $N^{-1/2}\sum_i\partial\ln\hat{f}_i/\partial\boldsymbol{\theta}|_{\boldsymbol{\theta}_0}$ 的矩阵倍数，它们既依赖于 $\partial\ln f_i/\partial\boldsymbol{\theta}$ 的可变性，又依赖于近似 $\hat{f}_i$ 的模拟误差。命题 12.1 表明，为使这个模拟误差消失，渐近采样次数必须以大于$\sqrt{N}$的速率随样本量而增大。

MSL 估计量的方差矩阵需要估计 $\mathbf{A}(\boldsymbol{\theta}_0)$。运用 5.5.2 节曾经定义的 BHHH 估计的模拟变形是最容易的。由于$\partial\ln f_i/\partial\boldsymbol{\theta}=(\partial f_i/\partial\boldsymbol{\theta})/f_i$，关于信息矩阵的 BHHH 估计是：

$$\hat{\mathbf{B}}=\frac{1}{N}\sum_{i=1}^{N}\frac{\partial f_i(\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}}{f_i(\hat{\boldsymbol{\theta}})}\ \frac{\partial f_i(\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}'}{f_i(\hat{\boldsymbol{\theta}})}$$

因为对于 f_i 以及$\partial f_i/\partial\boldsymbol{\theta}$ 来说，不存在闭形式解，故不能计算这个表达式。因此，用式(12.17)定义的模拟器 $\hat{f}_i$ 代替 f_i，得到渐近方差的模拟估计：

$$\hat{\mathrm{V}}[\hat{\boldsymbol{\theta}}_{\mathrm{MSL}}]=\left(\sum_{i=1}^{N}\left(\frac{\sum_{s=1}^{S}\partial\tilde{f}_i^s(\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}}{\sum_{s=1}^{S}\tilde{f}_i^s(\hat{\boldsymbol{\theta}})}\ \frac{\sum_{s=1}^{S}\partial\tilde{f}_i^s(\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}'}{\sum_{s=1}^{S}\tilde{f}_i^s(\hat{\boldsymbol{\theta}})}\right)\right)^{-1} \qquad (12.21)$$

其中 $\tilde{f}_i^s(\hat{\boldsymbol{\theta}})=\tilde{f}(y_i|\mathbf{x}_i,u_i^s,\hat{\boldsymbol{\theta}}_{\mathrm{MSL}})$。方差矩阵的可供选择的估计能通过类似于 5.5.2 节中定义的三明治方法估计海赛矩阵。

一个重要的实际问题是模拟次数。当样本量增大时，人们可增加模拟次数，但 S 的大小或其绝对值依然不确定。比如说，如果利用 2 400 次模拟与利用 2 600 次模拟进行估计时差异很小，那么我们把这看成 2 400 次模拟是一个足够多次数的象征。假定现在样本量增大 4 倍。我们应增加多少模拟次数呢？命题 12.1 表明，应该使 S 增大超过 2 倍，即大于 4 800，因而，$\sqrt{N}/S$ 比率趋于零而递减。然而，注意到，在此情况下，比如说当 $S=2\,400$ 且 $N=6\,400$，$\sqrt{N}/S$ 等于 1/30，我们不能确定这是否充分接近于 0。因此，对于人们是否做了充足多次的模拟问题，很难给出一个解答。许多应用者均依赖于点估计收敛的大致指示变量，即非正式地建立在检查 $\mathrm{L}_N(\boldsymbol{\theta})$的梯度基础上。一种选择 S 的基于检验的正式方法是由哈吉瓦斯利奥(Hajivassiliou，2000)给出并加以探讨的。

12.4.4 *调整渐近偏倚的 MSL*

当模拟次数 $S<\infty$时，MSL 估计量是非一致的或渐近有偏的。即使模拟器 $\hat{f}_i$ 关于 f_i 是无偏的，由于取自然对数的结果，$\ln\hat{f}_i$ 关于 $\ln f_i$ 是有偏的，所以便产生了这种偏倚。因而，对于有限 S 来说，$N^{-1}\ln(\hat{\mathrm{L}}_N(\boldsymbol{\theta}))$与 $N^{-1}\ln\mathrm{L}_N(\boldsymbol{\theta})$具有不同的概率极限。因为不能设 $S=\infty$，并且令 S 很大时，其计算量花费也很大，这就激励

了对可供选择的基于模拟估计量的探索研究。

一种明显的方法是，寻找关于对数密度 $\ln f_i$ 而不是 f_i 的无偏模拟器，但在实际应用中这样做行不通。相反，在本节，我们阐述 MSL 的校正偏倚形式，并在下一节阐述一种可供选择的比 MSL 稍欠有效的估计量，对于有限 S 来说是一致的。

古里耶克斯和蒙福特(Gouriéroux and Monfort，1991)已经给出 MSL 估计量的一种有偏表达式。对于固定 S 来说，MSL 估计量的非一致性源于下述事实：$\ln \hat{f}$ 是 $\ln f$ 的非一致估计量。减少非一致性的方法是，使用调整偏倚的对数似然函数。写成：

$$\ln \hat{f} = \ln[f+(\hat{f}-f)]$$

一旦在 $\ln f$ 附近取二阶泰勒级数展开式，得到：

$$\ln \hat{f} \simeq \ln f+\frac{\hat{f}-f}{f}-\frac{1}{2}\frac{(\hat{f}-f)^2}{f^2}$$

针对 $\mathbf{u}$ 的密度进行积分，并求解 $\ln f$，得到：

$$\ln f \simeq \mathrm{E}_{\mathbf{u}}[\ln \hat{f}]+\frac{1}{2}\frac{\mathrm{E}_{\mathbf{u}}[(\hat{f}-f)^2]}{f^2} \tag{12.22}$$

假定 $\hat{f}$ 是无偏模拟器，则 $\mathrm{E}_{\mathbf{u}}[\hat{f}]=f$。很明显，该表达式使得具有很小方差的模拟器 $\hat{f}$ 有较小偏倚。

校正偏倚估计量使用建立在式(12.22)右边项基础上的调整对数似然。对于模拟器(12.17)来说，$\hat{f}$ 等于 $S^{-1}\sum_s \tilde{f}^s$，而 $\mathrm{E}_{\mathbf{u}}[(\hat{f}-f)^2]$ 等于 $S^{-1}\sum_s \mathbf{E}_{\mathbf{u}}[(\tilde{f}^s-f)^2]$。已知对于不同 s 采样是独立的，后者能由 $S^{-1}\sum_s(\tilde{f}^s-\hat{f})^2$ 来逼近。那么，由式(12.22)得到，一阶渐近的校正偏倚的 MSL(**first-order asymptotic bias-corrected MSL**)估计量 $\hat{\boldsymbol{\theta}}_{\mathrm{BCMSL}}$，它对：

$$\ln \hat{\mathrm{L}}_{B,N}(\boldsymbol{\theta})=\sum_{i=1}^{N}\left[\ln \hat{f}(y_i\,|\,\mathbf{x}_i,\mathbf{u}_{iS},\boldsymbol{\theta})+\frac{1}{2S}\frac{\sum_{s=1}^{S}[\tilde{f}(y_i,\mathbf{x}_i,u_i^s,\boldsymbol{\theta})-\hat{f}(y_i\,|\,\mathbf{x}_i,\mathbf{u}_{iS},\boldsymbol{\theta})]^2}{\hat{f}(y_i\,|\,\mathbf{x}_i,\mathbf{u}_{iS},\boldsymbol{\theta})^2}\right]$$

求极大值，其中，$\hat{f}(y_i,\mathbf{x}_i,\mathbf{u}_{iS},\boldsymbol{\theta})=S^{-1}\sum_s \tilde{f}(y_i,\mathbf{x}_i,\mathbf{u}_i^s,\boldsymbol{\theta})$。由于偏倚是很小的假设可能并不总成立，故这种缩减偏倚方法的效果将会随着情况不同而变化。

12.4.5 *不可观测异质性例子*

假定 $y_i \sim \mathcal{N}[\theta_i,1]$，其中，纯量参数 θ_i 随不同个体而变化，即 $\theta_i=\theta+u_i$，而 u_i 表示非可观测的异质性，u_i 被假定成为服从已知分布。以 u 为条件的 y 的密度正是：

$$f(y\,|\,u,\theta)=\frac{1}{\sqrt{2\pi}}\exp\{-(y-\theta-u)^2/2\} \tag{12.23}$$

不过，对 θ 进行推断需要建立在 y 的边缘密度基础上(也就是说，关于 u 的边缘密度)，这需要积分去掉 u。此处，假定 u 具有密度：

$$g(u)=e^{-u}\exp(-e^{-u}) \tag{12.24}$$

为了简单起见，具有非零均值的斜分布并不依赖于未知参数。

由于边缘密度 $f(y|\theta)$ 等于$\int f(y|\theta,u)g(u)du$，没有闭形式解，所以极大似然估计行不通。相反，利用式(12.16)直接模拟器，我们运用MSL估计量，因而 $\hat{\theta}_{MSL}$ 对：

$$\ln \hat{L}_N(\theta)=\frac{1}{N}\sum_{i=1}^{N}\ln\left(\frac{1}{S}\sum_{s=1}^{S}\frac{1}{\sqrt{2\pi}}\exp\{-(y_i-\theta-u_i^s)^2/2\}\right) \tag{12.25}$$

求极大值，其中 u_i^s，$s=1,\cdots,S$ 表示从式(12.24)的极值密度 $g(u_i)$ 中采样，MSL估计量 $\hat{\theta}_{MSL}$是对一阶条件：

$$\frac{\partial \ln \hat{L}_N(\theta)}{\partial\theta}=\frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{s=1}^{S}(y_i-\theta-u_i^s)\exp\{-(y_i-\theta-u_i^s)^2/2\}}{\sum_{s=1}^{S}\exp\{-(y_i-\theta-u_i^s)^2/2\}}=0 \tag{12.26}$$

求解，并经过某种简化而得到的。θ 没有闭形式解，但可用一些标准迭代方法计算 $\hat{\theta}_{MSL}$。

除通常样本量 $N\to\infty$ 以外，MSL估计量的一致性要求采样数量 $S\to\infty$，因此，该方法潜在地是密集计算的。于是，像往常一样，MSL估计量服从渐近正态分布，其渐近方差最容易利用BHHH估计量(12.21)进行估计，从而得到：

$$\hat{V}[\hat{\theta}_{MSL}]=\left(\sum_{i=1}^{N}\left[\frac{\sum_{s=1}^{S}(y-\hat{\theta}_{MSL}-u_i^s)\exp\{-(y_i-\hat{\theta}_{MSL}-u_i^s)^2/2\}}{\sum_{s=1}^{S}\exp\{-(y_i-\hat{\theta}_{MSL}-u_i^s)^2/2\}}\right]^2\right)^{-1} \tag{12.27}$$

该估计量是完全有效的。

为了简述方便，我们考察满足 $\theta=1$ 的模型(12.23)与模型(12.24)所生成的样本量 $N=100$ 的样本$\{y_1,\cdots,y_{100}\}$。表12.2给出当采样数量 S 增大时的估计值。对于小 S 来说，MSL估计量是非一致的，对于 $S=10\ 000$ 来说，尽管估计的标准误差跳跃得相当大，但是估计量 $\hat{\theta}_{MSL}$稳定下来。当 S 增大时，模拟对数似然变小，但最终稳定下来。这种变小是人们所期望的，该模拟器关于 $f(y|\theta)$是无偏的，但关于 $\ln f(y|\theta)$ 却有向上偏倚，因为自然对数函数是全局凹的，由詹森不等式知，$\ln \mathrm{E}[\hat{f}(y|\theta)]>\mathrm{E}[\ln \hat{f}(y|\theta)]$；参见附录A(A.8节)。

表 12.2 极大模拟似然估计的例子

模拟次数	$S=1$	$S=10$	$S=100$	$S=1\ 000$	$S=10\ 000$
MLS估计值 $\hat{\theta}$	1.041 6	1.059 4	1.177 5	1.184 5	1.182 8
标准误差	(0.096 8)	(0.109 3)	(0.145 3)	(0.144 8)	(0.009 1)
$\ln\hat{L}(\hat{\theta})$	−136.31	−174.38	−190.44	−192.43	192.35

12.5 基于矩模拟估计

当目标函数不存在闭形式表达式时，模拟估计方法除了能被推广到MLE情况之外，还被推广到估计量上。此外，在一些情况下，对每个观测值仅仅使用很少模拟就可获得一致参数估计值是可能的，尽管这样做会损失有效性。

12.5.1 模拟 m 估计量

考察具有目标函数：

$$Q_N(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} q(y_i, \mathbf{x}_i, \boldsymbol{\theta})$$

的 m 估计量(参见 5.2.2 节)。极大似然是 $q(y,\mathbf{x},\boldsymbol{\theta})=\ln f(y|\mathbf{x},\boldsymbol{\theta})$的特殊情况。

假定 $q(\cdot)$不存在闭形式表达式,但可利用模拟估计。于是,模拟 m 估计量(**simulated m-estimator**)是对

$$\hat{Q}_N(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \hat{q}(y_i, \mathbf{x}_i, \mathbf{u}_{iS}, \boldsymbol{\theta}) \tag{12.28}$$

求极小值,其中类似于 12.4.1 节,$\hat{q}_i$ 表示建立在适当分布 S 个采样 $\mathbf{u}_i^s$ 的向量 $\mathbf{u}_{iS}$ 基础上 q_i 的估计值,$s=1,\cdots,S$。通常,$\hat{q}_i(\cdot)=S^{-1}\sum_i\tilde{q}(y_i|\mathbf{x}_i,\boldsymbol{\theta},\mathbf{u}_i^s)$,其中,$\mathbf{u}_i^s$ 表示第 s 次采样。

如果 m 估计量是一致的,并另外满足:

$$\operatorname{plim}\hat{Q}_N(\boldsymbol{\theta})=\operatorname{plim}Q_N(\boldsymbol{\theta}) \tag{12.29}$$

那么模拟 m 估计量将是一致的,因为由 5.3 节知,初始 m 估计量一致性的必要条件是 $\operatorname{plim}Q_N(\boldsymbol{\theta})$在 $\boldsymbol{\theta}=\boldsymbol{\theta}_0$ 处被极大化。此处,第一个 plim 是关于所有随机变量的,包括模拟采样 $\mathbf{u}_{iS}$,而第二个 plim 则不依赖于 $\mathbf{u}_{iS}$。

若模拟器使得当 $S\to\infty$时,$\hat{q}_i - q_i \xrightarrow{p} 0$,则条件(12.29)得到满足,从而 $N^{-1}\sum_i\hat{q}_i - N^{-1}\sum_i q_i \xrightarrow{p} 0$。如同 12.4 节一样,$S$ 随样本量增大,因此 $\sqrt{N}/S\to 0$,模拟 m 估计量应该与 m 估计量具有相同的极限分布。这需要许多次模拟。

12.5.2 减少模拟次数

现在,假定模拟器 $\hat{q}_i$ 不仅是一致的而且是无偏的。于是,通过应用大数定律,并且为了简单起见,除不用模拟采样以外,还不用随机变量,故 $\operatorname{plim}\hat{Q}_N(\boldsymbol{\theta})=\lim N^{-1}\sum_i \mathrm{E}_{\mathbf{u}_{iS}}[\hat{q}_i]=\lim N^{-1}\sum_i q_i=\operatorname{plim}Q_N(\boldsymbol{\theta})$,并且条件(12.29)得到满足。因而,倘若 $\mathrm{E}_{\mathbf{u}_{iS}}[\hat{q}_i]=q_i$,模拟 m 估计量只会与对每一个观测值 $\mathbf{u}_i$ 采样的情况相一致。

不幸的是,这一结果很难实施,因为在应用中几乎极少会找到 q_i 的无偏模拟器。例如,对于 ML 估计来说,找到密度 f_i 的无偏模拟器是可能的,但要找到 $\ln f_i$ 的无偏模拟器是不可能的。类似地,对于 NLS 估计来说,找到条件均值的无偏估计量是可能的,但是要找到误差平方的无偏模拟器是不可能的,这会涉及条件均值的平方。

可是,在一些情况下,当估计量是矩方法或 GMM 估计量而不是 m 估计量时,就可实现这一结果。

12.5.3 模拟矩方法

假定理论产生一个条件矩条件:

$$\mathrm{E}[m(y_i,\mathbf{x}_i,\boldsymbol{\theta}_0)|\mathbf{x}_i]=0 \tag{12.30}$$

为了简单起见，$m(\cdot)$表示纯量。设 $\mathbf{w}_i$ 表示工具，可能作为 $\mathbf{x}_i$ 与 $\boldsymbol{\theta}_0$ 的一个函数，满足：

$$\mathrm{E}[\mathbf{w}_i m(y_i, \mathbf{x}_i, \boldsymbol{\theta}_0)] = 0 \tag{12.31}$$

矩方法估计量 $\hat{\boldsymbol{\theta}}_{\mathrm{MM}}$（参见第 6 章 6.3.1 节）是对：

$$Q_N(\theta) = \left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{w}_i m(y_i, \mathbf{x}_i, \boldsymbol{\theta})\right]'\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{w}_i m(y_i, \mathbf{x}_i, \boldsymbol{\theta})\right] \tag{12.32}$$

求极小值，其中为了简单起见，假定是恰好识别情况：$\dim[\mathbf{w}_i] = \dim[\boldsymbol{\theta}]$。可把该结果推广到过度识别的情况，只是记号显得更繁琐，这是因为需要引进加权矩阵，并通过 GMM 加以估计。

矩方法估计量是一致的且具有正态分布，其方差矩阵部分地依赖于对工具 $\mathbf{w}_i$ 的选取。一个例子就是非线性回归，其中，$m(y, \mathbf{x}, \boldsymbol{\theta}) = y - \mathrm{E}[y|\mathbf{x}]$表示误差项，而条件均值 $\mathrm{E}[y|\mathbf{x}]$是 $\mathbf{x}$ 与 $\boldsymbol{\theta}$ 的一个设定函数。若误差是同方差的，则对工具的最佳选取是 $\mathbf{w} = \partial \mathrm{E}[y|\mathbf{x}]/\partial \boldsymbol{\theta}|_{\boldsymbol{\theta}_0}$，从而矩方法估计量与那些 *NLS* 估计量具有相同的一阶条件。

现在，假定 $m(y, \mathbf{x}, \boldsymbol{\theta})$没有闭形式表达式。例如，非线性回归模型可能缺少条件均值的闭形式表达式。然而，$m(y, \mathbf{x}, \boldsymbol{\theta})$是一个积分：

$$m(y_i, \mathbf{x}_i, \boldsymbol{\theta}) = \int h(y_i, \mathbf{x}_i, \mathbf{u}_i, \boldsymbol{\theta})]g(\mathbf{u}_i)d\mathbf{u}_i \tag{12.33}$$

对于某些 $h(\cdot)$与 $g(\cdot)$来说，它没有闭形式解。这时，不再可能有矩方法估计量。

模拟矩方法（**method of simulated moments**，记为 MSM）估计量 $\hat{\boldsymbol{\theta}}_{\mathrm{MSM}}$，它对

$$\hat{Q}_N(\boldsymbol{\theta}) = \left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{w}_i \hat{m}(y_i, \mathbf{x}_i, \mathbf{u}_{iS}, \boldsymbol{\theta})\right]'\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{w}_i \hat{m}(y_i, \mathbf{x}_i, \mathbf{u}_{iS}, \boldsymbol{\theta})\right] \tag{12.34}$$

求极小值，其中，$\hat{m}(y_i, \mathbf{x}_i, \mathbf{u}_{iS}, \boldsymbol{\theta})$表示关于 $m(y_i, \mathbf{x}_i, \boldsymbol{\theta})$的无偏模拟器（**unbiased simulator**），满足条件：

$$\mathrm{E}[\hat{m}(y_i, \mathbf{x}_i, \mathbf{u}_{iS}, \boldsymbol{\theta})] = m(y_i, \mathbf{x}_i, \boldsymbol{\theta}) \tag{12.35}$$

并且 $\mathbf{u}_{iS}$表示从边缘密度 $g(\mathbf{u}_i)$中得到的 S 个采样，而 $S \geqslant 1$。下面，将给出 m_i 与无偏模拟器 $\hat{m}_i$ 的一些例子。

12.5.4 *MSM 估计量的分布*

MSM 估计量是由麦克法登（McFadden，1989）提出的，他已经证明该估计量具有下述性质。

命题 12.2（MSM 估计量的分布）［麦克法登（McFadden，1989）］ 假定下述条件：

(i) 数据来自数据生成过程的简单随机样本，其中，$m(y, \mathbf{x}, \boldsymbol{\theta}_0)$具有如同式（12.30）的零条件期望，并且 $\mathbf{w}_i m(y, \mathbf{x}, \boldsymbol{\theta}_0)$具有如同式（12.32）的零无条件期望，同时一些条件得到满足，以使对式（12.32）求极小值的 MM 估计量是一致的且渐近正态的。

(ii) 函数 $m(y,\mathbf{x},\boldsymbol{\theta}_0)$ 是由式(12.33)定义的，并可利用满足式(12.35)的无偏模拟器 $\hat{m}(y,\mathbf{x},\boldsymbol{\theta}_0)$ 来进行估计。

于是，对于固定 S 来说，当 $N\to\infty$ 时，对式(12.34)求极小值的模拟矩方法估计量(**method of simulated moments estimator**)是一致且渐近正态的，并服从极限正态分布，满足：

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{\mathrm{MSM}}-\boldsymbol{\theta}_0)\xrightarrow{d}\mathcal{N}[\mathbf{0},\ \mathbf{A}^{-1}(\boldsymbol{\theta}_0)\mathbf{B}(\boldsymbol{\theta}_0)\mathbf{A}^{-1}(\boldsymbol{\theta}_0)'] \tag{12.36}$$

其中：

$$\mathbf{A}(\boldsymbol{\theta}_0)=\operatorname{plim}\frac{1}{N}\sum_{i=1}^{N}\mathbf{w}_i\left.\frac{\partial m_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}_0} \tag{12.37}$$

并且：

$$\mathbf{B}(\boldsymbol{\theta}_0)=\operatorname{plim}\frac{1}{N}\sum_{i=1}^{N}\mathbf{w}_i\mathrm{V}[\hat{m}_i(\boldsymbol{\theta}_0)]\mathbf{w}_i' \tag{12.38}$$

这里，方差 V[•]既是关于给定 x_i 时 y_i 的条件分布，又是关于给定式(12.35)之后采样 $\mathbf{u}_{iS}$ 的。

在给出该命题推导之前，我们注意到下述内容。第一，即使 $S=1$，MSM 估计量具有显著的成为一致的性质。第二，对于有限的 S 来说，会损失有效性。$\hat{\boldsymbol{\theta}}_{\mathrm{MM}}$的方差矩阵与 $\hat{\boldsymbol{\theta}}_{\mathrm{MSM}}$的相同，只是对于 MM 估计而言，式(12.38)中的 $\mathrm{V}[\hat{m}_i]$用较小的 $\mathrm{V}[m_i]$代替。第三，当 $S\to\infty$时，由模拟引起的有效性损失将会消失，从而 $\mathrm{V}[\hat{m}_i]=\mathrm{V}[m_i]$。第四，就 MM 估计而言，如果工具 $\mathbf{w}$ 选择不好，那么与其他一些估计量相比，尽管 $S\to\infty$，但 MSM 估计量仍可能是无效的。

MSM 估计量的一致性要求，对于已知式(12.34)与式(12.32)中的 $\hat{Q}_N(\boldsymbol{\theta})$与 $Q_N(\boldsymbol{\theta})$来说，条件(12.29)要得到满足。由大数定律得：

$$\operatorname{plim}\frac{1}{N}\sum_{i=1}^{N}\mathbf{w}_i\hat{m}_i=\operatorname{plim}N^{-1}\sum_{i=1}^{N}\mathbf{w}_i\mathrm{E}_{\mathbf{u}_{iS}}[\hat{m}_i]$$

其中，第一个 plim 是关于所有随机变量的，而第二个 plim 是关于除模拟采样 $\mathbf{u}$ 以外的所有随机变量。这里，$\mathrm{E}_{\mathbf{u}_{iS}}[\hat{m}_i]=m_i$，因为 $\hat{m}_i$ 是一个无偏模拟器，因此有：

$$\operatorname{plim}\frac{1}{N}\sum_{i=1}^{N}\mathbf{w}_i\hat{m}_i=\operatorname{plim}N^{-1}\sum_{i=1}^{N}\mathbf{w}_i m_i$$

这同样也蕴含，$\operatorname{plim}\hat{Q}_N(\boldsymbol{\theta})=\operatorname{plim}Q_N(\boldsymbol{\theta})$。所以，倘若 $\boldsymbol{\theta}_0$ 使 $\operatorname{plim}Q_N(\boldsymbol{\theta})$极大化，这是最初 MM 估计量成为一致的所必需的，则 $\hat{\boldsymbol{\theta}}_{\mathrm{MSM}}$是一致的。

对于极限分布来说，对 $\hat{Q}_N(\boldsymbol{\theta})$ 求关于 $\boldsymbol{\theta}$ 的微分，得到：

$$\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{w}_i\frac{\partial\hat{m}_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right)'\frac{1}{N}\sum_{i=1}^{N}\mathbf{w}_i\hat{m}_i(\hat{\boldsymbol{\theta}})=\mathbf{0}$$

第一个矩阵是满秩的方阵，因而，$\hat{\boldsymbol{\theta}}_{\mathrm{MSM}}$等价地满足一阶条件：

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{w}_i\hat{m}_i(\hat{\boldsymbol{\theta}})=\mathbf{0}$$

其中，$\hat{m}_i(\boldsymbol{\theta})=\hat{m}_i(y_i,\mathbf{x}_i,\mathbf{u}_{iS},\boldsymbol{\theta})$。利用通常在 $\boldsymbol{\theta}_0$ 附近的准确一阶泰勒级数展开式，得出：

$$\sum_{i=1}^{N}\mathbf{w}_i\hat{m}_i(\boldsymbol{\theta}_0)+\sum_{i=1}^{N}\mathbf{w}_i\left.\frac{\partial\hat{m}_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}^*}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)=\mathbf{0}$$

从而有：

$$\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)=-\left(N^{-1}\sum_{i=1}^{N}\mathbf{w}_i\left.\frac{\partial\hat{m}_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}^*}\right)^{-1}N^{-1/2}\sum_{i=1}^{N}\mathbf{w}_i\hat{m}_i(\boldsymbol{\theta}_0)$$

现在 $\mathrm{E}_{\mathbf{u}}[\partial\hat{m}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}]=\partial\mathrm{E}_{\mathbf{u}}[\hat{m}(\boldsymbol{\theta})]/\partial\boldsymbol{\theta}=\partial m(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$，因此，右边的第一个矩阵收敛到由命题 12.2 给出的 $\mathbf{A}(\boldsymbol{\theta}_0)$。右边的第二项服从极限正态分布，其均值为 0 且方差矩阵为：

$$\mathbf{B}(\boldsymbol{\theta}_0)=\mathrm{plim}\frac{1}{N}\sum_{i=1}^{N}\mathbf{w}_i\mathrm{V}[\hat{m}_i(\boldsymbol{\theta}_0)]\mathbf{w}_i'$$

如同命题 12.2 一样，其中，$\mathrm{V}[\hat{m}_i(\boldsymbol{\theta}_0)]$表示关于 $\mathbf{u}_{iS}$和给定 $\mathbf{x}_i$ 时 y_i 分布的方差。

由于 $\mathbf{u}_{iS}$与 y_i 是独立的，从而：

$$\begin{aligned}\mathrm{V}_{y,\mathbf{u}}[\hat{m}(\boldsymbol{\theta}_0)]&=\mathrm{V}_y[\mathrm{E}_{\mathbf{u}}[\hat{m}(\boldsymbol{\theta}_0)]]+\mathrm{E}_y[\mathrm{V}_{\mathbf{u}}[\hat{m}(\boldsymbol{\theta}_0)]]\\&=\mathrm{V}_y[m(\boldsymbol{\theta}_0)]+\mathrm{E}_y[\mathrm{V}_{\mathbf{u}}[\hat{m}(\boldsymbol{\theta}_0)]]\end{aligned}$$

代入上式后，就会得出命题 12.2 给出的 $\mathbf{B}(\boldsymbol{\theta}_0)$的更详细的定义。

由于出现 $\mathrm{E}_y[\mathrm{V}_{\mathbf{u}}[\hat{m}(\boldsymbol{\theta}_0)]]$项，模拟会使 MSM 估计量的方差增大，当$S\to 0$ 时，$\mathrm{E}_y[\mathrm{V}_{\mathbf{u}}[\hat{m}(\boldsymbol{\theta}_0)]]$趋于 0。在特殊情况下，模拟器是频率模拟器，可以证明，$\mathrm{V}_{y,\mathbf{u}}[\hat{m}(\boldsymbol{\theta}_0)]=(1+1/S)\mathrm{V}_y[m(\boldsymbol{\theta}_0)]$，故利用频率模拟器的模拟效果会使 MM 估计量的方差扩大到$(1+(1/S))$!。

12.5.5 *在 MSM 与 MSL 之间选择*

应用者将会对 MSL 与 MSM 的优缺点进行权衡。已知 MSM 对于小的 S 是一致的，并且进一步地为了保证对 MLE 具有一个良好近似，确保人们具有充分大的 S 集合是困难的，和 MSM 相比，为什么 MSL 更受欢迎呢?

首先，注意到 MSL 原则上简单易行，并且直接实施。给定参数假设，对于 MLE 方法来说，观测值的最优加权是内在的。与之相比，类似于 GMM，MSM 要求我们对权数（或工具变量）函数与残差的乘积进行计算，而这些成分可能是相关的。例如，GMM 估计量的数值（不含模拟的）不稳定性已由奥尔顿吉和西格尔（Altonji and Segal, 1996）（参见 6.3.5 节）证明过。类似地，格韦克、基恩和朗克尔（Geweke, Keane, and Runkle, 1997）以及麦克法登和鲁德（Mcfadden and Ruud, 1994）都曾经提供 MSM 估计量不稳定性的证据。不过，尽管简单性有助于 MSL，但是与确保应用时有足够多模拟次数相联系的一些问题不应受到低估。

12.5.6 *不可观测的异质性例子*

我们回到 12.4.5 节的例子上。于是，$y_i\sim\mathcal{N}[\theta+u_i,\ 1]$，其中，$u_i$ 具有式(12.24)给出的密度 $g(u_i)$。由于 $\mathrm{E}[y_i-\theta-u_i]=0$，所以利用矩方法估计量来估计 θ，该估计量是

$$\frac{1}{N}\sum_{i=1}^{N}(y_i-\theta-\mathrm{E}[u_i])=0\qquad(12.39)$$

的解，从而得到 $\hat{\theta}_{MM}=\bar{y}-\mathrm{E}[\bar{u}]$。假定 $\mathrm{E}[\bar{u}]$是未知的，却能用 MSM 估计量 $\hat{\theta}_{MSM}$进行求解：

$$\frac{1}{N}\sum_{i=1}^{N}\left(y_i-\theta-\frac{1}{S}\sum_{s=1}^{S}u_i^s\right)=0 \tag{12.40}$$

其中，u_i^s 表示从极值分布中得到的 iid 随机采样。

对估计方程(12.40)求解，得到：

$$\hat{\theta}_{MSM}=\bar{y}-\bar{\bar{u}} \tag{12.41}$$

其中，$\bar{\bar{u}}=(NS)^{-1}\sum_i\sum_s u_i^s$ 表示既对 N 又对 S 求平均值。不过，更一般地讲，计算 MSM 估计量可能需要迭代法。

很容易获得 $\hat{\theta}_{MSM}$的方差。由构造知，u 的模拟采样互相之间是独立的，并具有原始数据 y 的形式，因此 $\mathrm{V}[\hat{\theta}_{MSM}]=\mathrm{V}[\bar{y}]+\mathrm{V}[\bar{\bar{u}}]$。现在，$\mathrm{V}[\bar{y}]=(\sigma_u^2+1)/N$。由于$\bar{\bar{u}}$表示 u 的 NS 个采样平均，故 $\mathrm{V}[\bar{\bar{u}}]=\sigma_u^2/NS$，由此可得：

$$\begin{aligned}\mathrm{V}[\hat{\theta}_{MSM}]&=\mathrm{V}[\bar{y}]+\mathrm{V}[\bar{\bar{u}}]\\&=\frac{\sigma_u^2+1}{N}+\frac{\sigma_u^2}{NS}\end{aligned} \tag{12.42}$$

这里用到了 $\hat{\sigma}_u^2=(NS)^{-1}\sum_{i=1}^{N}\sum_{s=1}^{S}(u_i^s-\bar{\bar{u}})^2$，从而得到一致估计。

考察来自满足 $\theta=1$ 的模型(12.24)所生成的样本量 $N=100$ 的样本$\{y_1,\cdots,y_{100}\}$。表 12.3 给出，当采样次数 $S\to\infty$时，MSM 估计量的情况。当模拟次数 S 增大时，MSM 估计量接近于矩方法估计，而且标准误差下降。

表 12.3 模拟矩估计方法的例子

模拟次数	$S=1$	$S=10$	$S=100$	$S=1\,000$	$S=\infty$ (MM)
MSM 估计值 $\hat{\theta}$	1.007 3	1.109 6	1.201 2	1.188 7	1.187 9
标准误差	(0.247 1)	(0.165 7)	(0.168 1)	(0.167 6)	(0.168 4)

12.6 间接推断

在本节，我们概述另一种基于模拟的针对模型进行估计的方法，有时当人们想要运用一种模型或相对简单地估计模型时，就要用到这种方法，甚至当基本数据生成过程被认为是更复杂且较难估计时。该方法存在几种变形与解释；参见古里耶克斯、蒙福特和雷诺尔特(Gouriéroux, Monfort, and Renault, 1993)、史密斯(Smith, 1993)、加伦特和陶亨(Gallant and Tauchen, 1996)。此方法有时还被称为矩匹配(**moment matching**)方法。本节的解释本质上沿着前面提及的第一类参考文献线索阐述。

假定在参数形式上被设定的数据生成过程用 pdf $f(\mathbf{y};\boldsymbol{\theta})$，$\boldsymbol{\theta}\in\mathcal{R}^q$ 表示，其参数相对很难估计出来。假定我们能设定含有数据生成过程 $f^a(\mathbf{y};\boldsymbol{\beta})$，$\boldsymbol{\beta}\in\mathcal{R}^r$ 的辅助模型(**auxiliary model**)，这很容易通过拟(有时，还被称为"伪")极大似然方法进行估计。由于下面将要进一步讨论的识别原因，假定 $\boldsymbol{\beta}$的维数并不小于 $\boldsymbol{\theta}$ 的维数，即

$r \geqslant q$。例如，辅助模型可以是对精确似然的一种近似，或者它可以是近似模型的一种精确似然。对于给定样本来说，设 $\hat{\beta}$ 表示 QML 估计值。然后，由 5.7 节讨论的结果，我们知道，$\hat{\beta}$ 通常是 $\boldsymbol{\theta}$ 的非一致估计量，同时在某些正则条件下，它依概率收敛到被称为伪真实(**pseudo-true**)值的值上，这是 $\boldsymbol{\theta}$ 的一个函数。把辅助模型的参数与数据生成过程的那些参数联系起来的函数被称为绑定函数(**binding function**)，记为 $\mathbf{h}(\boldsymbol{\theta})$。该函数的解析形式可能是已知的，也可能是未知的。因此，想要得到 $\boldsymbol{\theta}=\mathbf{h}^{-1}(\boldsymbol{\beta})$ 或 $\hat{\boldsymbol{\theta}}=\mathbf{h}^{-1}(\hat{\boldsymbol{\beta}})$，可能并不总是可行的。

间接推断的方法可用于获得一个更优的 QML 估计量，它的渐近偏倚要小于 $\hat{\boldsymbol{\beta}}$ 的渐近偏倚。其思想是，在 $f(\mathbf{y};\boldsymbol{\theta})$ 下使用模型通过模拟伪观测值 $\mathbf{y}^{(s)}$ 生成，并在 $f^a(\mathbf{y}^{(s)};\boldsymbol{\beta})$ 下，运用辅助回归来估计 $\hat{\boldsymbol{\beta}}^{(s)}$，其中，$s$ 表示第 s 次模拟。间接估计量通过对

$$\hat{\boldsymbol{\theta}}=\arg\min_{\theta}(\hat{\boldsymbol{\beta}}^{(s)}-\hat{\boldsymbol{\beta}})'\boldsymbol{\Omega}(\hat{\boldsymbol{\beta}}^{(s)}-\hat{\boldsymbol{\beta}}) \tag{12.43}$$

求解加以定义，其中，$\boldsymbol{\Omega}$ 表示给定的对称正定矩阵。该估计量类似于 6.7 节曾经考虑的最小距离估计量。也就是说，我们可连续不断地生成伪观测值，并对建立在伪观测值基础上的辅助模型的参数进行估计。这种迭代连续不断地进行，直到式(12.43)的二次形式被极小化。一个非常重要的关键问题是，生成伪随机观测值 $\mathbf{y}^{(s)}$ 的种子(**seed**)要保持不变，因此，伪观测值对于不同模拟的变异归因于 $\hat{\boldsymbol{\beta}}^{(s)}$ 中的变异。

在进一步讨论之前，考察一个简单却包括非线性数据生成过程与线性辅助模型的特定例子。其动机是，辅助模型应该是容易估计的，同时数据生成过程也应是容易估计的。

设数据生成过程具有如下形式：

$$y_i=\exp(\mathbf{x}_i'\boldsymbol{\gamma})+u_i \tag{12.44}$$
$$u_i \sim \mathcal{N}[0,\sigma^2]$$

设其辅助模型是下述形式：

$$y_i=\mathbf{x}_i'\boldsymbol{\beta}+\varepsilon_i \tag{12.45}$$
$$\varepsilon_i \sim \mathcal{N}[0,\sigma_\varepsilon^2]$$

注意下述解释：

$$\frac{\partial \mathrm{E}[y|\mathbf{x}]}{\partial \mathbf{x}}=\boldsymbol{\beta} \quad (\text{在辅助回归模型下})$$

$$\frac{\partial \ln \mathrm{E}[y|\mathbf{x}]}{\partial \mathbf{x}}=\frac{\partial \mathrm{E}[y|\mathbf{x}]}{\partial \mathbf{x}}\times\frac{1}{\mathrm{E}[y|\mathbf{x}]}=\boldsymbol{\gamma} \quad (\text{在该数据生成过程下})$$

因此，绑定函数是 $\boldsymbol{\gamma}\mathrm{E}[y_i|\mathbf{x}]=\boldsymbol{\beta}$，或者 $\boldsymbol{\gamma}=(\mathrm{E}[y_i|\mathbf{x}])^{-1}\boldsymbol{\beta}$。注意，$\dim(\boldsymbol{\beta})$ 等于 $\dim[\boldsymbol{\gamma}]$。

已知数据($\mathbf{x}_i$, y_i, $i=1,\cdots,N$)与最小二乘法估计量 $\hat{\boldsymbol{\beta}}$，并已知 N 维伪随机采样，记为 $u^{(0)}$，利用：

$$y_i^{(1)}=\exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}})+u_i^{(0)}$$

生成 $y_i^{(1)}(i=1,\cdots,N)$，并且获得一个修正估计量 $\hat{\beta}^{(1)}=(\sum \mathbf{x}_i\mathbf{x}_i')^{-1}\sum \mathbf{x}_i y_i^{(1)}$，这同样可用于生成其他的伪观测值集合。一旦使 $\mathbf{u}^{(0)}$ 固定，整个模拟过程是一个反复过程，一直到$(\hat{\beta}^{(s)}-\hat{\beta})'\mathbf{\Omega}(\hat{\beta}^{(s)}-\hat{\beta})$接近于人们期望精确度的常值为止。在此情况下，有理由设 $\mathbf{\Omega}$ 等于单位矩阵或者等于 $\mathbf{X}'\mathbf{X}$，而后一种选择意味着，来自辅助模型的预测是对目标的建模。所得到的 γ 的估计就是间接估计量。

在其他一些应用中，$\dim(\beta)$将大于 $\dim(\boldsymbol{\theta})$，因而没有唯一的 $\boldsymbol{\theta}$ 值可被利用。实际上，在缺乏解析绑定函数的情况下，即使两个维数都是一样的，也不能重新获得 $\boldsymbol{\theta}$。于是，人们对辅助模型参数的最佳间接估计感到满意。

为了理解间接估计量与矩匹配(**moment matching**)之间的联系，设 $\mathbf{\Omega}=\mathbf{X}'\mathbf{X}$；那么$(\hat{\beta}^{(s)}-\hat{\beta})'\mathbf{X}'\mathbf{X}(\hat{\beta}^{(s)}-\hat{\beta})=(\hat{\beta}^{(s)}\mathbf{X}-\hat{\beta}\mathbf{X})'(\hat{\beta}^{(s)}\mathbf{X}-\hat{\beta}\mathbf{X})$，这表明间接估计量是与分布的一阶矩相“匹配”的。当人们还想要匹配二阶矩时，通过其他参数比如方差参数对向量 β进行扩大。因此，如果人们愿意的话，就能匹配多阶矩。

在正则条件下，间接估计量是一致的且渐近正态的。对于其他详细内容，读者参见前面引述的研究。

12.7 模拟器

如同 12.3.2 节一样，计算：

$$I=\mathrm{E}[h(x)]=\int h(x)g(x)dx \tag{12.46}$$

其中，为了简单起见，x 经常表示纯量。正如 12.3 节，此处 x 经常用于表示想要积分去掉的变量，而在应用部分中，当 $\mathbf{x}$ 表示回归元时，就用 $\mathbf{u}$ 表示想要积分去掉的变量。

模拟器是计算 I 的一种方法。除由式(12.14)给出的直接蒙特卡罗积分法以外，还有许多方法可以应用。原则上，模拟器应是一个无偏的模拟器，并是光滑的，因此使用标准的迭代梯度法。虽然情况如此，但在实证研究时，对感兴趣的模型进行估计时，所耗计算时间则是一个难以克服的障碍。我们对众多巧妙方法中的几个加以阐述，这几种方法对于任何给定的模拟采样来说，通过减少像直接蒙特卡罗积分法这类粗糙方法的模拟方差，来加以模拟。一个更完整的综述，已由格韦克和基恩(Geweke and Keane，2001)给出。

12.7.1 *频数模拟器*

我们以一个例子开始，即能用于某些离散模型的频数模拟器。这将突出模拟时产生的几个新困难。

假定函数 $h(x)$表示指示函数，即当 $x\in A$ 时，$h(x)$取值 1，否则为 0。于是，想要计算：

$$I=\int \mathbf{1}(x\in A)g(x)dx$$

由直接蒙特卡罗积分法，得出估计值：

$$\hat{I}_{\text{FREQ}} = \frac{1}{S}\sum_{s=1}^{S}\mathbf{1}(x^s \in \mathbf{A})$$

其中，x^s 表示从 $g(x)$ 中得到的 S 个采样，$s=1,\cdots,S$。这称为频数模拟器（**frequency simulator**），因为它是通过 x^s 的 S 个采样落入 A 之中的相对频数来进行估计的。

一个重要的潜在应用是——由在模拟方法方面的许多经济计量学文献所激发——12.2.2 节引入的多项式离散选择模型。对于三种可供选择的模型来说，选择由式(12.3)给出的第一种可供选择的概率为 p_1，对二变量正态分布的正象限积分。从而，频数模拟器 $\hat{p}_1$ 是从满足 $u_1^s \geqslant 0$ 且 $u_2^s \geqslant 0$ 的二变量正态分布得到的采样 (u_1^s, u_2^s) 比例。

频数模拟器有几个局限性。第一，它既不是可微的，又不是关于参数 $\boldsymbol{\theta}$ 连续的，$\boldsymbol{\theta}$ 出现在 $\mathbf{1}(x\in A)$ 与/或 $g(x)$ 之中。

因此，$\boldsymbol{\theta}$ 小的变化都会导致相同采样数落入正象限中。鉴于这种原因，麦克法登（McFadden, 1989）以及帕克斯和波拉德（Pakes and Polland, 1989）曾经提出涵盖这类非光滑模拟器的更一般渐近理论。不过，在实际应用中，一种最好的方法是，运用可供选择的关于参数是可微的光滑模拟器（**smooth simulators**），因为这允许利用通常的梯度法进行计算。

第二，如果仅有一小部分 $x\in A$，那么模拟器是非常无效的。例如，对于含有 $p_1=0.001$ 的离散选择模型来说，甚至拥有 10 000 个采样，估计值 $\hat{p}_1$ 将含有相当大的噪声。更一般地讲，如果采样 x 的概率在 $h(x)$ 相对大的范围内是很小的，那么具有连续 $h(x)$ 的直接蒙特卡罗计算式(12.46)会产生类似问题。

第三，即使模型利用 $0<I<1$，并对估计模型来说，这个条件是必需的，此模拟器在边界上可能出现问题，给出估计值 $\hat{I}=0$ 或 $\hat{I}=1$。

12.7.2 *重要抽样*

重要抽样模拟器（**important sampling simulators**）是将积分式(12.46)重新写成：

$$I = \int\left(\frac{h(x)g(x)}{p(x)}\right)p(x)dx \tag{12.47}$$
$$= \int w(x)p(x)dx$$

其中，$p(x)$ 表示选取的密度函数，以使：(a) 很容易从 $p(x)$ 中获得采样；(b) $p(x)$ 与最初的积分定义域具有相同的支集；(c) $w(x)=h(x)g(x)/p(x)$ 很容易计算，并且是有界的，同时具有有限方差。然后，我们运用基于式(12.47)而不是式(12.46)的直接蒙特卡罗积分法估计：

$$\hat{I}_{\text{IS}} = \frac{1}{S}\sum_{s=1}^{S}w(x^s) \tag{12.48}$$

其中，x^s 均是从 $p(x)$ 而不是从 $g(x)$ 采样，$s=1,\cdots,S$。重要的抽样术语是由 $w(x)$ 决定样本空间中不同点的权数或者“重要性”而得来的。在贝叶斯模拟文献中，重要抽样已被运用许多年，它是由克洛克和范迪克（Kloek and van Dijk, 1978）引入贝叶斯经济计量学中作为计算后验分布的一种方法。13.4 节将进一步讨论这个

内容。

已知来自 $p(x)$ 的独立采样，重要抽样器 $\hat{I}_{IS}$ 具有方差 $S^{-1}V_p[w(x)]$。很明显，当 $w(x)$ 在整个积分范围上是一个常值时，方差达到最小值，从而 $V_p[w(x)]$ 为 0。这通过令 $w(x)=E_g[h(x)]$ 来完成，进而 $p(x)=h(x)g(x)/E_g[h(x)]$ 是积分为 1 的密度。不幸的是，在理论上，这种理想的重要抽样估计行不通，因为 $E_g[h(x)]$ 是未知的。不过，尤其是如果选取 $p(x)$ 以使 $w(x)$ 是相当平坦的，那么它表示运用重要抽样的潜在好处。

即使重要抽样使方差增大，在实际应用时会出现这种情况，它却具有其他方面的吸引力。若 $w(x)$ 在估计参数方面是光滑的，则它产生光滑抽样器。此外，倘若从 $g(x)$ 采样很困难，就可运用该方法，正如当 x 是相关时，随机变量经常遇到的情况。

对于多项式 probit 离散选择模型来说，一种流行的重要抽样器是 GHK 抽样器(**GHK simulator**)，归功于格韦克(Geweke, 1992)、哈吉斯利奥和麦克法登(Hajivassiliou and McFadden, 1994)，还有基恩(Keane, 1994)。这种模拟器会递归地截取多变量正态 pdf，以使采样被限制在正象限上。与频数模拟器相比，该模拟器的优点是，它是光滑的，需要较少的具有很小概率的可供选择的采样，并不可能出现边界问题。

12.7.3 *用对偶加速缩减方差*

前面方法都假定从适当分布诸如 $g(x)$ 中独立采样，或者如果使用重要抽样，利用将在 12.8 节详述的一些方法，从 $p(x)$ 中独立采样。

相反，方差减少(**variance reduction**)方法使用相关采样，因为这些能缩减模拟器的方差。一个重要例子是对偶抽样(**antithetic sampling**)，它运用负相关的采样。里普利(Ripley, 1987,第 129～132 页)、格韦克(Geweke, 1988)以及哈吉瓦斯利奥(Hajivassiliou, 2000)都对这种方法进行了讨论，而格韦克(Geweke, 1995)曾经综述这个方法以及其他几种缩减方差的方法。

假定我们想要计算式(12.46)的积分 I，其中，x 被假定成具有零均值且对称密度为 $g(x)$。建立在从 $g(x)$ 中得到的 $2S$ 个模拟 iid 采样基础上的直接蒙特卡罗积分法估计是：

$$\hat{h}_{2S}(x) = \frac{1}{2S}\sum_{s=1}^{2S} h(x^s)$$

而且，已知 $2S$ 个采样的独立性，则有方差：

$$V[\hat{h}_{2S}(x)] = \frac{1}{2S}V[h(x)]$$

对偶抽样(**antithetic sampling**)运用仅仅建立在 S 个 iid 采样基础上的可选择估计值是：

$$\hat{h}_{A,S}(x) = \frac{1}{S}\sum_{s=1}^{S} \frac{1}{2}(h(x^s) + h(-x^s)) \tag{12.49}$$

这表示 $h(x)$ 在 x^s 与 $-x^s$ 上计算值的平均。序对(x^s, $-x^s$)称为对偶序对(**anti-**

thetic pair)，同时由于假定 x 是均值为 0 的对称分布，所以得到一个无偏的估计值。不过，当均值为 μ 时，$(x^s, 2\mu-x^s)$ 就是一个对偶序对。已知 x^s 的 S 个独立采样，$\hat{h}_{A,S}(x)$ 方差是：

$$\begin{aligned} V[\hat{h}_{A,S}(x)] &= \frac{1}{S^2}\sum_{s=1}^{S}\frac{1}{4}(V[h(x^s)]+2\mathrm{Cov}[h(x^s), h(-x^s)]+V[h(-x^s)]) \\ &= \frac{1}{2S}(V[h(x)]+\mathrm{Cov}[h(x), h(-x)]) \end{aligned}$$

因此，当协方差项为负时，对偶抽样将比常规的 iid 抽样更为有效，从而 $\hat{h}_{A,S}(x)$ 的方差比 $\hat{h}_{2S}(x)$ 的要小。通过改变采样的符号，然后再用采样，以此尝试缩减模拟器的负相关。当函数是线性的时候，若非线性不太严重，可确定有负相关。可是，通常人们不能确信有效性提高将会实现。例如，如果 $h(\cdot)$ 关于 0 是对称的，那么 $\mathrm{Cov}[h(x), h(-x)]=V[h(x)]$。

可将对偶抽样推广到对称密度 $g(x)$ 上。假定 x 可利用稍后 12.8.2 节给出的逆变换方法进行采样。比如说，人们从均匀分布[0, 1]中采样 u 来生成对偶变换 $(1-u)$，然后运用逆变换方法从选择的分布进行采样，因此 $x_1=G^{-1}(u)$ 且 $x_2=G^{-1}(1-u)$，其中，$G(\cdot)$ 表示 x 的已知 cdf。于是，当：

$$\mathrm{Cov}[h(G^{-1}(u)), h(G^{-1}(1-u))]=\mathrm{Cov}[f(u), f(1-u)]<0$$

(x_1, x_2) 就形成有对偶序对，并出现方差缩减，其中，$f(u)$ 表示复合函数 $h(G^{-1}(u))$。如果 $f(\cdot)$ 表示单调函数，那么方差就减少了[罗伯特和卡塞拉(Robert and Cesella, 1999, 第 112 页)]。不过，该函数的这一性质很难得以验证。进一步地，此讨论仅仅用于逆变换方法，而在实际应用中其他一些方法可用于伪随机数生成(参见 12.8 节)。因此，在特定应用中，要达到预先验证有效性提高(增益)的一些条件是很困难的。

尽管在更复杂设置的情况下，戏剧性提高有效性可能在一些特殊情况下不会出现，但在许多情况下，提高有效性是值得做的。对偶抽样还能用于加速重要抽样[丹尼尔森和理查德(Danielsson and Richard, 1993)]。

可将对偶抽样推广到多变量抽取。考察二变量采样 (x, y)，其密度关于 $(0, 0)$ 对称。在这种情况下，符号反转要先用于逐个元素，然后形成一个序对。因而，对偶四元组是由 $((x^s, y^s)(-x^s, y^s), (x^s, -y^s), (-x^s, -y^s))$ 构成的。对于 m 维采样来说，同样的思想对所有多元组不断重复进行。

12.7.4 *用准随机序列计算*

第二种缩减方差的方法是，用准随机数[1] (**quasi-random numbers**)代替伪随机数，其目的是提供更好的样本空间运用系统模拟采样。该方法的潜在局限性是，随机性要求应用大数定律与中心极限定律为基于模拟的方法提供证据。

准蒙特卡罗方法在积分定义域中使用非随机点代替使用 S 个伪随机点。一个重要例子是，由普雷斯等人(Press et al., 1993)概括的霍尔顿序列(**Halton sequences**)，

---

[1] 又称为拟随机数。——译者注

并由布哈特(Bhat, 2001)与特雷恩(Train, 2003)引入经济计量学文献中。

霍尔顿序列拥有两个人们期望的性质。第一,利用它们设计成给出抽样分布定义域的相当均匀的范围。就每个观测值的更均匀散布采样而言,相对于那些随机采样计算来说,模拟概率会很少随不同观测值而变化。这类似于特定网格上对积分进行确定计算。第二,就霍尔顿序列而言,一个观测值的采样会填满由先前观测值留下的空白区。因此,模拟概率对不同的观测值是负相关的。如同对偶变量情况一样,这种负相关缩减了模拟函数的方差。在适当正则条件下,可以证明,与收敛速率为 N^{-1} 的伪随机序列相比,利用准随机序列的积分误差是阶数为 $N^{-1/2}$ 的形式[布哈特(Bhat, 2001)]。

霍尔顿序列最好通过例子加以阐述。假定作为模拟的函数依赖于单个随机变量。其起点是一个素数。建立在素数 2 基础上的霍尔顿序列是如下构造的。把单位区间(0,1)分成两部分。分割点 1/2 成为霍尔顿序列的第 1 个元素。接下来,把每个部分分成两个部分。分割点 1/4 与 3/4 成为该序列的随后两个元素。对这四个部分的每一个都分成两部分,从而连续不断地获得序列{1/2, 1/4, 3/4, 1/8, 3/8, …}。类似地,建立在素数 3 基础上的序列是{1/3, 2/3, 1/9, 2/9, 4/9, …}。建立在非素数上的霍尔顿序列不是唯一的,因为非素数的霍尔顿序列以同样方式对单位空间进行分割作为构造非素数的素数。

每个序列长度由观测值个数 N 与模拟采样 S 次数来决定。对于不同素数的霍尔顿序列来说,当前面一些元素具有相关趋势时,人们将放弃该序列的前几个(比如说,20 个)[例如, 参见特雷恩(Train, 2003)]。因此,人们能通过生成长度为 $N\times S+20$ 的霍尔顿序列开始,然后放弃每个序列的前 20 个元素。对于每个序列的每一个元素来说,计算累积正态分布的逆。所得到的值是,源自抽样分布的霍尔顿采样(**Halton draws**)。

准随机数采样的一个重要优点是,采样被用于设计成比在伪随机数情况下更均匀的方式涵盖随机数的样本空间。实际上,由图 12.1 已经看到这点。在该图上,第 2 个图形表明,利用霍尔顿序列从所构造的二变量正态分布中得到的采样。其余三个图形表明,从同一分布中得到的伪随机数。在前者情况下,显然样本空间范围显得更为均匀。

在一维或者多维情况下,使用霍尔顿采样的基于模拟估计的例子与更为深入全面的讨论,以及给人留下深刻印象的该种方法的相对有效性证据,参见特雷恩(Train, 2003, 第 9 章)。对于服从正态分布的随机参数的多项式 logit 模型来说,该方法会很好地起作用(参见 15.7 节)。

12.8 随机变量采样方法

前面模拟器要求对随机变量进行采样。在本节,我们概述从密度中进行这类采样的一些方法,在 12.7 节将密度记为 $g(x)$或 $p(x)$,而在本节则将其记为 $f(x)$。通常,从均匀分布或标准正态分布(在大部分流行软件里,这是可能的)中获得采样就足够了,因为这些能形成不同于均匀或正态分布中得到的采样。

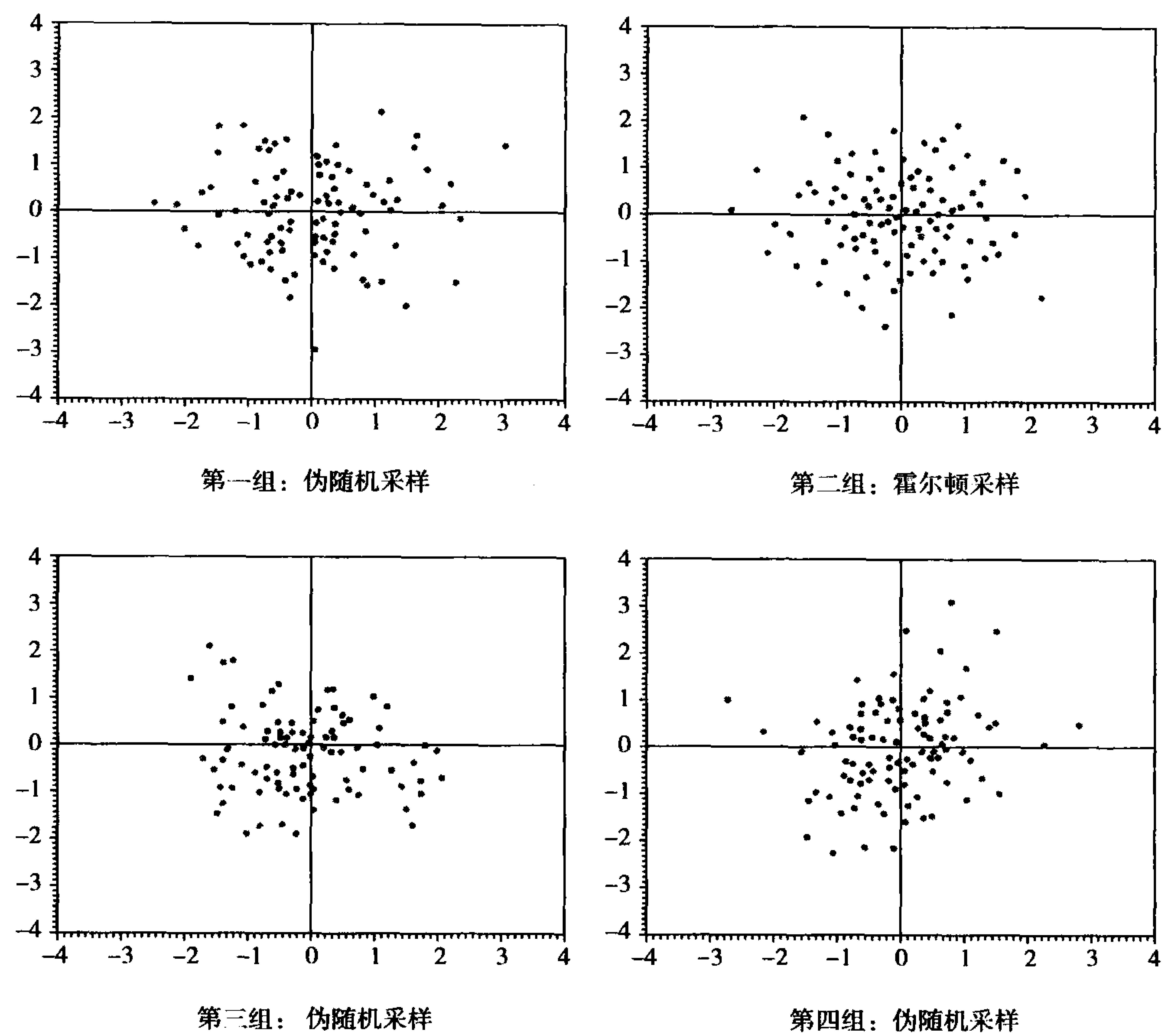


图 12.1 霍尔顿采样(第二组)与伪随机采样的比较

如果采样是用于基于模拟的估计，为了防止"颤动"，那么所有出自均匀或标准正态的采样都应该在任何估计之前完成，迭代法收敛失败的原因在于，在每一次迭代时由新的采样所产生的噪声。例如，若 $x \sim \mathcal{N}[\mu, \sigma^2]$ 且 μ 与 σ 的估计随着迭代而变化，则我们做出 $z \sim \mathcal{N}[0, 1]$ 的 NS 个最初采样，然后对于不同迭代利用 z 的最初采样，重新计算 $x=\mu+\sigma z$。

本节提供关于生成随机变量的某些标准方法的基本讨论。对于更高等的广泛研究内容有许多好的著作，包括由布拉德利、福克斯和施拉格(Bradler, Fox, and Schrage, 1983)，达格珀纳(Dagpunar, 1988)，德夫罗尔(Devroye, 1986)，以及里普利(Ripley, 1987)的那些文献。

在阐述方法之前，注意，随机数生成术语是一个矛盾形容法(oxymoron)。一种更准确的描述，可通过伪随机数(**pseudo-random numbers**)给出。这些生成元的基本特征是，它们使用确定性装置生成可模仿来自某个目标分布实现性质的很长的一串数。特定的目标分布将依赖于背景内容，只是对本书来说，均匀分布、正态分布、指数分布、伽玛分布、logistic 分布以及泊松分布都是标准的。一连串过程则是通过提供种子数(**seed**)开始的。在某个有限却很大的次数之后，对数自身周而复始的重复过程便生成了。也就是说，计算机算法将会准确地生成以给定种子开

始的相同数。一个好的随机数生成元是，可不用再次循环且不用任何内置相依性而生成一长串数。在选择生成元中，一个重要的考虑是，在合理计算成本下生成分布是否密切地模仿了目标分布的性质。

12.8.1 *伪随机均匀数生成元*

伪随机均匀数（**pseudo-random uniform numbers**）是利用模仿均匀随机数序列的统计性质的确定性序列而构造的。一个好的生成元具有很长的周期，接近于均匀分布，同时生成独立的采样。重要的是拥有一个好的生成元，实际上作为来自任何分布的伪随机数，都能通过对均匀伪随机数进行变换而获得[布拉德利等人(Bradley et al.，1983，第 24 页)]。

一个标准的生成元是以方程：

$$X_j=(kX_{j-1}+c)\ \mathrm{mod}\ m$$

开始，其中，当 a 被 b 除时，模数映射 a mod b 形成了余数。从而，产生 0 与 1 之间的整数序列，然后，获得均匀随机变量作为 $R_j=X_j/m$[里普利(Ripley，1987，第 20 页)]。对于 X_0 的一个值——称之为种子数(**seed**)，需要引进生成元。所生成的均匀随机序列都是确定性的，倘若分析是以该种子的同样数值进行重复，这便像应采样同样的数那样使复制成为可行的。如果计算是利用 32 位整数完成的，那么算术最大周期性近似为 $2^{31}\simeq2.1\times10^{9}$。不过，容易选到不好的 X_0、k 以及 c，所以周期性比这要更小些。诸如普雷斯等人(Press et al.，1993)所撰写的书，都提及潜在陷阱的问题。

12.8.2 *非均匀变量*

来自许多其他分布，包括正态分布自身的一些随机变量，通常都建立在均匀随机数的最初采样上。四种普遍运用的方法是：(1)逆变换；(2)变换；(3)筛选法；(4)混合与合成。

逆变换

设 $F(x)$ 表示连续随机变量 x 的 cdf，即：

$$F(x)=\Pr[X\leqslant x]$$

已知均匀变量的一个采样 r，$0\leqslant r\leqslant1$，逆变换(**inverse transformation**)：

$$x=F^{-1}(r)$$

就给出 x 的唯一值，因为 F 是连续且单调递增的。

例如，单位指数的 cdf 是 $1-e^{-x}$。求解 $r=1-e^{-x}$，得出 $x=-\ln(1-r)$。当我们从均匀[0，1]中进行采样且得到 0.64 时，$x=-\ln(1-0.64)$。图 12.2 画出 X 的 cdf，并显示出这种方法从图形上看是如何起作用的。纵坐标轴上的任何点在高度 r 处选取，而其横轴上的对应值可通过画出长方形来获得。这就是逆变换。

尤其是，如果 $F(\cdot)$ 的解析形式已知，并且 x 是连续随机变量，那么这一方法很容易使用。若没有闭形式表达式可利用，则该方法经常仍是可行的，尽管在计算上代价会更高一些，因为标准分布的逆 cdf 经常在程序中作为函数而得以利用。

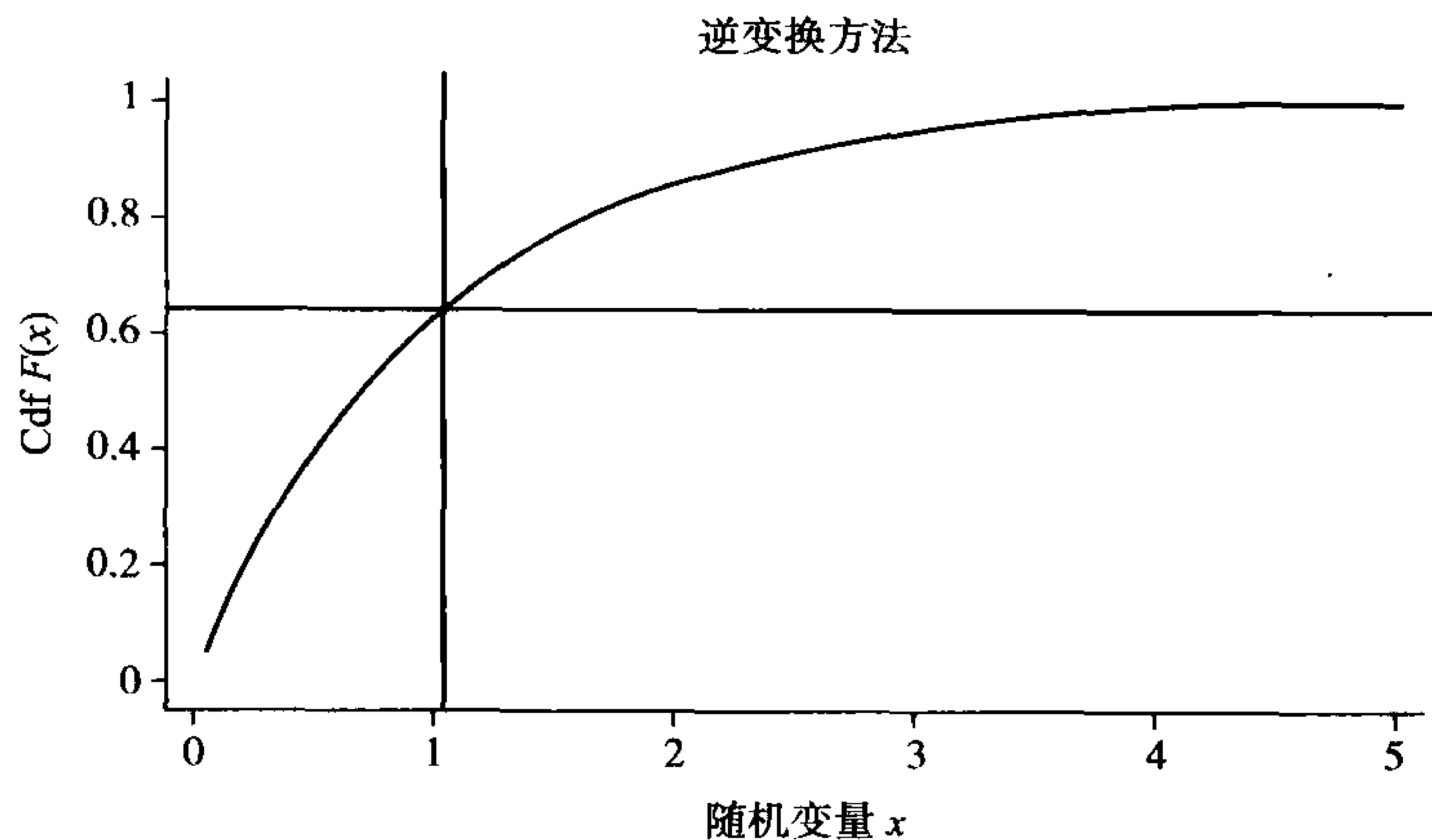


图 12.2 从单位指数中进行采样的逆变换。随机均匀采样为 0.64 [因而，$F(x)=1-\exp(-x)=0.64$]，得出 $x=1.02$。

这个方法能被推广到 cdf 是阶梯函数的离散变量上，例如，当 x 取整数值时，均匀采样 $r=0.312$ 会得出 $x=j$ 的采样，其中，整数 j 使得 $F(j-1)<0.312$ 且 $F(j)>0.312$。

生成正态随机变量的标准方法是博克斯—米勒(Box - Muller)方法。这要运用逆变换方法，把其应用于两个独立的正态变量联合分布而不是单个变量上。特别地，若 r_1 与 r_2 都是 iid 且均匀的，则 $x_1=\sqrt{-2\ln r_1}\cos(2\pi r_2)$ 与 $x_2=\sqrt{-2\ln r_1}\times\sin(2\pi r_2)$ 都是 iid 的并服从 $\mathcal{N}[0,1]$。

变换

在一些情况下，具有人们期望密度的随机变量，通过对很容易从其分布中抽取的那些随机变量，进行适当的变换(**transformation**)而获得。然后，通过运用这个相同变换得到随机变量。

这样的变换方法是，一种明显的从基于正态分布中获得采样的方法。一些例子包括，平方标准正态变量可获得含有中心卡方分布的随机变量，一旦添加 r 个独立的标准正态变量的平方值，会产生具有 r 个自由度的卡方变量，同时计算独立卡方的均值平方，可产生 F 分布的随机变量。变换方法并没有被限制在基于正态分布上。

筛选法

假定我们想要从密度 $f(x)$ 中进行采样，这很困难，不过，存在对于所有 x，对某一有限常值 k，在 $f(x)\leqslant kg(x)$ 的意义上，涵盖 $f(x)$ 的另外一个密度 $g(x)$，这已画在图 12.3 之中，其中，粗线用来拟合包络线 $kg(x)$。

筛选法(**accept-reject method**)是从 $g(x)$ 而不是从 $f(x)$ 获得采样。当：

$$r\leqslant\frac{f(x)}{kg(x)}$$

就接收采样，$x=r$，其中，r 表示从均匀分布得到的采样。若上述条件不满足，就拒

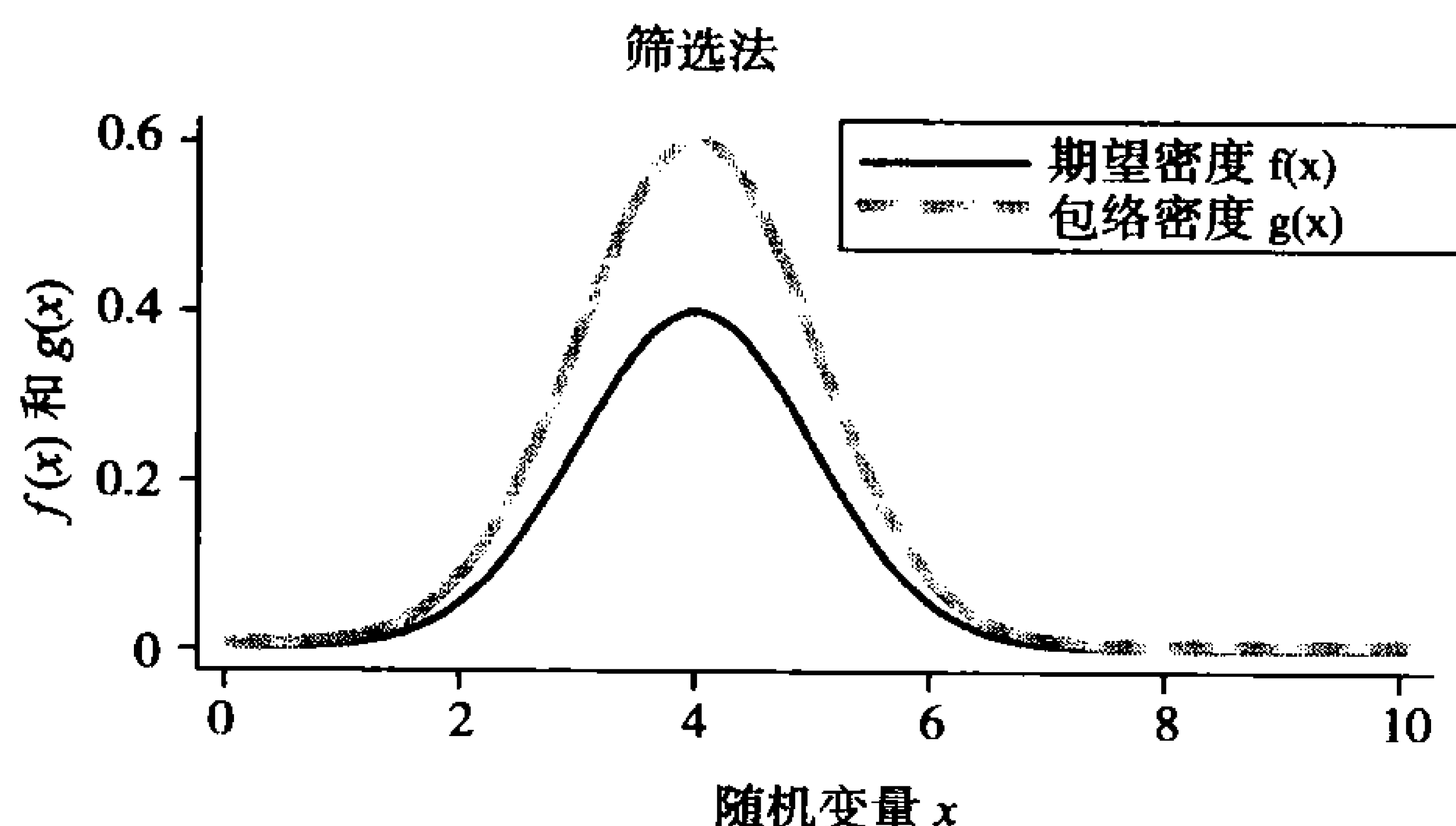


图 12.3 从密度 $g(x)$中采样的筛选法，其中，$kg(x)$包络了人们期望的密度 $f(x)$。

绝采样，并进一步地进行采样，一直到条件得以满足。该方法的吸引力依赖于从 $g(x)$中很容易获得采样而不是 $f(x)$ 采样。其局限性在于平均采样将以概率 $1/k$ 被接收，所以当 k 很大时，就要求有众多采样。

为了理解这一方法是如何起作用的，设 Y 表示通过筛选法生成的随机变量，X 表示具有密度 $g(x)$的随机变量，而 U 表示从均匀分布得到的采样。于是，Y 具有 cdf：

$$\begin{aligned}
\Pr[Y \leqslant y] &= \Pr[X \leqslant y \mid U \leqslant f(x)/kg(x)] \\
&= \frac{\Pr[X \leqslant y,\ U \leqslant f(x)/kg(x)]}{\Pr[U \leqslant f(x)/kg(x)]} \\
&= \frac{\int_{-\infty}^{y}\int_{0}^{f(x)/kg(x)} du\, g(x)dx}{\int_{-\infty}^{\infty}\int_{0}^{f(x)/kg(x)} du\, g(x)dx} \\
&= \frac{\int_{-\infty}^{y}[f(x)/kg(x)]g(x)dx}{\int_{-\infty}^{\infty}[f(x)/kg(x)]g(x)dx} \\
&= \frac{\int_{-\infty}^{y}[f(x)/k]dx}{\int_{-\infty}^{\infty}[f(x)/k]dx} \\
&= \int_{-\infty}^{y} f(x)dx
\end{aligned}$$

如同所期望的，这是对应于密度 $f(x)$的 cdf。

合成

有时，密度 $f(x)$能被表述成来自混合分布或复合分布的形式，满足：

$$f(x) = \int g(x \mid \varepsilon)h(\varepsilon)d\varepsilon$$

于是，源自 $f(x)$的采样能通过首先从密度 $h(\varepsilon)$获得采样，然后从条件密度 $g(x \mid \varepsilon)$ 进行 x 的采样而获得。

举一个例子，考察从均值为 λ 而方差为 $\lambda(1+\alpha\lambda)$的负二项分布进行采样，其

中，λ 与 α 都是给定常值。这里我们使用将负二项分布看成泊松—伽玛的混合的事实（参见第 20 章）。首先，从均值为 1 且方差为 α 的伽玛采样 ε，这通过指数变换来完成。其次，从均值为 $\lambda\varepsilon$ 的泊松分布进行采样，ε 是由前面一节给定的。

当 $h(\varepsilon)$ 表示在 C 个点处具有质量 p_j 的离散分布，$j=1,\cdots,C$，上面积分步骤可由求和来代替。因而，有：

$$f(x)=\sum_{j=1}^{C}p_jg(x|\varepsilon=\varepsilon_j)$$

然后，为了从 $f(x)$ 获得 S 个采样，我们从每一个 $g(x|\varepsilon=\varepsilon_j)$ 中采样 Sp_j 个观测值，并通过混合采样“合成”所需要的 S 个值的样本。

一些标准生成元

附录 B 中的表描述了几种标准的连续情况与离散情况的伪随机数生成。它们是建立在下述假设基础上的，即 $r,r_1,r_2,\cdots$都是独立均匀[0, 1]中随机变量 $R,R_1,R_2,\cdots$的值。注意到，可能存在各种不同方法来生成相应的随机变量，我们仅仅列出这些方法中的一两个。

12.8.3 *多元分布*

从多元分布[1]（multivariate distributions）中采样，通常比从一元分布中采样更为复杂。例如，诸如逆变换与变换等方法不再可应用。对于一些多元分布来说，使用混合方法或者合成方法，因为一些多元分布就是混合分布。

相当一般的方法就是，吉布斯抽样以及其他的马尔可夫链蒙特卡罗方法。这些内容将推迟到 13.5 节，因为它们广泛地应用于贝叶斯分析中，运用复杂多元分布。正如将要解释的那样，利用吉布斯抽样器所做的采样，可能显示出呈现某种相关的趋势，此事实将减少模拟器的有效性。

这里，将注意力限制在多元正态分布上。于是，采样很容易通过一元标准正态采样的变换获得。特别地，假定我们想要从 q 维正态分布中进行采样，因而 $\mathbf{x}\sim\mathcal{N}(\mathbf{0},\Sigma)$。这通过建立在正定 Σ 具有乔列斯基分解（**Choleski decomposition**）：

$$\boldsymbol{\Sigma}=\mathbf{LL}'$$

基础上的变换完成，其中，$\mathbf{L}$ 表示下三角矩阵。例如，对于 $q=2$ 来说，乔列斯基分解是：

$$\begin{bmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{12} & \sigma_{22}\end{bmatrix}=\begin{bmatrix}l_{11} & 0\\ l_{21} & l_{22}\end{bmatrix}\begin{bmatrix}l_{11} & l_{21}\\ 0 & l_{22}\end{bmatrix}$$

从而，得到三个方程 $l_{11}^2=\sigma_{11}$，$l_{11}l_{21}=\sigma_{12}$，而且 $l_{21}^2+l_{22}^2=\sigma_{22}$，利用它们求解 l_{11}、l_{21} 以及 l_{22}。给定 q 维向量 $\boldsymbol{\varepsilon}$，其元素具有标准正态分布，容易验证，若 $\boldsymbol{\varepsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})$，则 $\mathbf{x}=\mathbf{L}\boldsymbol{\varepsilon}$，正态线性组合服从分布 $\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})$。特别地，$\mathrm{E}[\mathbf{L}\boldsymbol{\varepsilon}]=\mathbf{0}$，而 $\mathrm{V}[\mathbf{L}\boldsymbol{\varepsilon}]=\mathrm{E}[\mathbf{L}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{L}']=\mathbf{LL}'=\boldsymbol{\Sigma}$。这个方法的关键是，正态的线性组合仍然是正态分布，此结果对于非正态分布不成立。

---

[1] 又称为多变量分布。——译者注

12.9 文献注释

普雷斯等人(Press et al.，1993)对求积分和蒙特卡罗积分提供了良好的起点，并给出进一步研究的参考文献，包括本章其他地方提及的某些文献。

有关基于模拟估计的经济计量学文献，强调多项式 probit 模型。不过，这类方法具有较广泛的应用性，同时在其他一些模型中实施起来更容易且更有成效，与多项式 probit 相比，其拟合时很少受到挑战。莱尔曼和曼斯基(Lerman and Manski，1981)使用模拟频率来估计选择概率，并发现要求众多采样。麦克法登(McFadden，1989)提出 MSM，同时阐述它的一致性与渐近正态性。帕克斯和波拉德(Pakes and Pollard，1989)已经提供既有 MSM 又有 MSL 的渐近理论相当一般性的研究。斯特恩(Stern，1997)的相对通俗易懂的综述则是一个优秀的开始点。吉利诺克斯和蒙福特(Gouriéroux and Monfort，1996)提供一种基本方法教科书式的研究。在后面特定几章将要讨论的模型背景下，许多其他参考文献更适合阅读。特别地，哈吉瓦斯利奥和鲁德(Hajivassiliou and Ruud，1994)强调包括多项式 probit 的截取正态模型，而特雷恩(Train，2003)曾经考察一系列离散选择模型，包括随机参数 logit。

习 题

12－1 通过蒙特卡罗来积分法估计 $I=\int t(x)g(x)dx$，已知和 $\hat{I}=N^{-1}\sum\times t(x_i)g(x_i)/p(x_i)$。其中，$x_i$ 表示从重要抽样分布 $p(x)$得到的采样。证明 $\text{plim}\,\hat{I}=I$。

12－2 对于 $f(\theta)=|\Sigma|^{-1/2}\left[1+\frac{1}{v}(\theta-\mu)'\Sigma^{-1}(\theta-\mu)\right]^{-(v+d)/2}$ 来说，考察 d 维积分 $\int_{R^d}f(\theta)d\theta$。被积函数是多元 t 密度的核，因此，正确解答是正规化常值的逆。

(a) 把这个积分作为蒙特卡罗平均 $S^{-1}\sum_{s=1}^{S}f(\theta^{(s)})/h(\theta^{(s)})$加以计算，$\theta^{(s)}\sim h(\theta)$，其中，重要密度 $h(\theta)$表示具有相同位置与标度 $f(\theta)$，却具有不同自由度参数的多元 t 密度。

(b) 当你变动 $h(\theta)$的自由度时，探索该平均的稳定性。通过变化 $h(\theta)$的位置与标度来增大 $f(\theta)$与 $h(\theta)$之间的错误匹配，并进一步加以探索。

12－3 对于 12.5.3 节的 MSM 估计量，假定模拟器是频数模拟器。

(a) 证明，$V_{y,u}[\hat{m}(\boldsymbol{\theta}_0)]=(1+1/S)V_y[m(\boldsymbol{\theta}_0)]$。

(b) 由此证明，利用频数模拟器的模拟效果会使矩方法估计量的方差膨胀。

(c) 当 $s=10$ 时，标准误差的有效性损失会是多大呢？

12－4 对于 12.5.6 节中的例子，考察作为$\sum_{i=1}^{N}\left[y_i-\frac{1}{S}\sum_{s=1}^{S}(\alpha+u_i^s)\right]=0$ 解的估计量 $\hat{\alpha}$。求此估计量及其方差的解析表达式。

12－5　(a) 写出从三维多元正态分布$\mathcal{N}[\mathbf{0}, \Sigma]$中采样伪随机样本的算法，其中满足$\sigma_{jj}=1$，$j=1,2,3$，同时协方差$\sigma_{12}=\sigma_{13}=\sigma_{23}=0.5$。采样 1 000 个实现的样本，并把估计的均值及方差与数据生成过程的那些均值及方差进行比较。

(b) 用具有 5 个自由度的学生 t 分布代替(a)中的三元正态分布，重复(a)部分内容。

12－6　利用 12.8.2 节给出的逆变换方法，写出从一元截尾正态密度$\mathcal{TN}_{[a, b]}[\mu, \sigma^2]$中进行采样的计算程序。这里$[a, b]$表示上截断点与下截断点。选取$\mu=1$，$\sigma^2=4$并且$a=3$，$b=4$。

12－7　考察标准二值 logit 回归模型(参见 14.3 节)。

(a) 写出对数似然函数。

(b) 当截距是从具有有限均值及方差的适当分布中采样得来的，引进随机截取假设。你对以这种方式引入的不可观测异质性会做出什么判断？如果 logit 模型是从含有极端值误差的随机效用模型中推导出来的，那么影响解释与推导的随机截距会怎样呢？[参见雷维尔特和特雷恩(Revelt and Train, 1998)。]

(c) 对随机截距提出一个适宜的分布假设；重新写出以不可观测异质性为条件的似然函数。然后，写出积分去掉的不可观测异质性的似然函数。

(d) 一步一步描述，如何用极大模拟似然估计方法估计这个模型。详细解释，如何计算未知参数的方差矩阵？如何决定你所使用的模拟次数？

(e) 考察模拟矩方法作为随机参数 logit 的 MSL 程序的可供选择方法。写出以不可观测异质性项为条件的矩条件。然后，概述此模型的 MSM 估计程序。

12－8　有些计算软件包允许你既可直接采样泊松伪随机数，又可直接采样伽玛伪随机数。而且，众所周知，负二项分布被推导成泊松随机变量与伽玛随机变量的混合(参见 20.4 节)。

(a) 写出利用混合方法采样负二项分布变量的程序。

(b) 用你的方法，对均值为 0.25 的泊松分布变量采样 10 000 次的样本。

(c) 从均值为 1 且方差为 α 的伽玛分布中采样相应样本，用 α 集合生成方差为 0.312 5 的负二项随机变量。

13 贝叶斯方法

13.1 引 论

本章介绍贝叶斯经济计量学的内容。自从泽尔纳(Zellner, 1971)与利默(Leamer, 1978)的书出版以来,贝叶斯回归分析以惊人的速度得到不断发展。常规数据分析应用也得到了巨大扩展,这在很大程度上得益于计算机硬件和软件技术方面革命性的进步。从这类重要发展来看,单独一章不足以合理应对该主题的众多方面。因此,本章对贝叶斯经济计量学的主要思想及发展提供一个大略的路线图,这是本章非常适宜的目标。尽管这是一个适宜目标,但仍有部分内容表现出相当的技术性。

与前面几章曾经阐述的似然或频率学派或者经典方法不同,贝叶斯方法需要对有关未知参数先验信念的概率模型进行设定。不论是在哲学上还是在实践上,研究者对这种步骤感到不尽如人意。在传统上看,这会涉及贝叶斯方法是主观的而不是客观的基础。可以证明,在大样本条件下,可忽略先验作用,并设定相对非信息的先验,而且可利用对先验敏感性的研究推断方法。因此,主观性变化并不总是像许多表述的那样严重。

在应用微观经济计量学中,特别是,当研究缺少解析形式易于处理的似然函数复杂模型时,贝叶斯方法将起到潜在而巨大的作用。第12章已介绍了面对这类情形的基于模拟方法,这些方法均潜在地存在问题,尤其是模拟似然法,因为它们通常要求利用充分大的模拟采样次数求函数的最大值,而模拟采样次数会随样本量增大以适当速率增加。即使拥有当今运算能力超强的计算机,对大样本与高维数模型进行分析,也需要求解难以克服的计算量。与之相比,贝叶斯方法并不要求最大值算法。贝叶斯方法灵活,足以产生不算极好的估计值,但在许多情况下,仍有效获得这种估计值。实际上,没有必要促使人们通过改变哲学上的信仰而运用贝叶斯方法,从实用主义考量,这类方法却有存在的必要。

以上评论并不意味着,贝叶斯方法没有比较深奥的理论基础和论证。贝叶斯方法具备这些方面。尤其是,值得提出三个特性。第一,贝叶斯方法能获得关注参数的整个后验分布,使用户潜在地根据决策理论准则去决定报告分布的矩及分位数。人们不要求各自独立的均值、中位数、分位数等一些估计量,因为后验分布都

包含它们。第二,以数据为条件的贝叶斯分析会得出准确的样本结果,消除了对有限样本进行修正或调整的要求。这一分布在大样本中接近于正态分布,先验的影响将会消失。第三,贝叶斯方法提供了选择模型的自然方法。

13.2 节介绍贝叶斯分析的基本概念与构成,以及贝叶斯估计量的重要性质。这些思想在 13.3 节以相对容易处理的线性回归模型加以阐明。更一般地,对于后验分布来说,不存在闭形式解。13.4 节阐述获得后验矩的数值估计的蒙特卡罗积分法,即著名的重要抽样。13.5 节详述马尔可夫链蒙特卡罗方法,包括著名的吉布斯抽样与梅特罗波利斯—黑斯廷斯算法,用于从(不易处理)后验分布中获得采样。这些方法的例子在 13.6 节给出。

此外,数据增广以及贝叶斯模型选择的专题,将在 13.7 节和 13.8 节阐述。

13.2 贝叶斯方法

在贝叶斯方法中,关于参数 $\boldsymbol{\theta}$ 值的不确定性是通过引入先验分布(**prior distribution**)的密度 $\pi(\boldsymbol{\theta})$ 而以显性方式得以建模的,这样命名是因为它没有考虑现有手头数据而加以设定。它用概率语言表述关于真实未知参数的主观信念。13.2.4 节将详细研究先验的设定。举一个例子,假定 θ 表示收入弹性,并根据经济模型或先前研究,认为 θ 以概率 0.95 位于 0.8 与 1.2 之间。那么,关于 θ 的先验信息就是 $\theta \sim \mathcal{N}[1,\ 0.1^2]$。

贝叶斯推断的其他构成部分是,样本联合密度或似然函数 $f(\mathbf{y}|\boldsymbol{\theta})$,在单方程情况下,$\mathbf{y}$ 表示 $N\times 1$ 维向量。为了记号简单起见,本节自始至终不用关于回归元的相依性。外生回归元将在 13.3 节引入,在此情况下,$f(\mathbf{y}|\boldsymbol{\theta})$ 变成 $f(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})$,贝叶斯分析是以回归元为条件的。还要注意到,在本章,$f(\cdot)$ 通常表示所有观测值的联合密度,而不是第 i 个观测值的密度。

若没有数据可利用,则我们拥有的全部就是先验信念。当数据是可观测的,经典方法是利用极大似然原理估计未知参数 $\boldsymbol{\theta}$。相反,贝叶斯方法是将样本的似然与先验结合起来,反映任何先验信息都应该得到探索的观点,尽管先验信息仅揭示出概率分布形式而已。这种过程被认为是,给定数据(似然)时对先验的一种修正。事实上,我们在将似然与先验结合后能够推导 $\boldsymbol{\theta}$ 的分布。所得到的分布称为后验分布(**posterior distribution**),同时它反映出研究者关于 $\boldsymbol{\theta}$ 的后验信念,也就是观测到数据之后的信念。

13.2.1 *贝叶斯定理*

提供后验分布的基本结果是贝叶斯定理(**Baye's Theorem**),有时还称为贝叶斯逆概率律(**inverse law of probability**),即:

$$f(\boldsymbol{\theta}|\mathbf{y})=\frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\mathbf{y})} \tag{13.1}$$

其中,$f(\mathbf{y})$ 表示 $\mathbf{y}$ 的边缘概率分布,正式地,定义:

$$f(y)=\int_{R(\theta)} f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{13.2}$$

其中，$R(\boldsymbol{\theta})$ 表示 $\pi(\boldsymbol{\theta})$ 的支集。获得该结果源于注意到，对于事件 A 与 B，条件概率为：

$$\begin{aligned}\Pr[A|B]&=\frac{\Pr[A\cap B]}{\Pr[B]}\\&=\frac{\Pr[B|A]\Pr[A]}{\Pr[B]}\end{aligned}$$

其中，第二个等式成立是因为 $\Pr[B|A]=\Pr[A\cap B]/\Pr[A]$。

由于式(13.1)中分母 $f(\mathbf{y})$ 不含有 $\boldsymbol{\theta}$，所以将 $p(\boldsymbol{\theta}|\mathbf{y})$ 更简单地写成正比于 pdf 与先验之积；因而：

$$p(\boldsymbol{\theta}|\mathbf{y})\propto \mathrm{L}(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \tag{13.3}$$

这可通过省略无关紧要的常值，得到简化后验的推导及表示式，省略常值稍后能重新获得，正如 13.2.2 节将阐明的那样。当密度函数不带正规化常值而被写出时，它称为密度核(**density kernel**)。

在许多情况下，式(13.1)或式(13.3)并不会产生后验密度的闭形式表达式。然而，不需要闭形式表达式，而后面几节将阐述用于获得对后验密度的基于模拟方法的良好数值近似。这些方法允许贝叶斯分析用于几乎任何的参数微观经济计量学应用。

运用关于后验密度的特定符号是普遍的，故将用 $p(\boldsymbol{\theta}|\mathbf{y})$ 代替 $f(\boldsymbol{\theta}|\mathbf{y})$。同理，最初的联合密度 $f(\mathbf{y}|\boldsymbol{\theta})$ 表示 $\mathrm{L}(\mathbf{y}|\boldsymbol{\theta})$ 的似然函数。此后，我们将把后验密度(**posterior density**)写成：

$$p(\boldsymbol{\theta}|\mathbf{y})\propto \mathrm{L}(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \tag{13.4}$$

这种表达式是贝叶斯方法的核心内容，该式强调了频率学派与贝叶斯方法之间的重要差异。在频率学派方法中，参数的真值是常值，但将参数估计值处理成随机变量。与之相比，在贝叶斯方法中，参数被处理成好像它是随机的。

13.2.2 *贝叶斯定理例子*

假定 $y\sim\mathcal{N}[\theta,\sigma^2]$，其中，$\sigma^2$ 已知，但纯量参数 θ 未知。已知随机样本$(y_1,\cdots,y_N)$，y 的联合密度是：

$$\begin{aligned}\mathrm{L}(\mathbf{y}|\theta)&=\prod_{i=1}^{N}(2\pi\sigma^2)^{-1/2}\exp\{-(y_i-\theta)^2/2\sigma^2\}\\&=(2\pi\sigma^2)^{-N/2}\exp\left\{-\sum_{i=1}^{N}(y_i-\theta)^2/2\sigma^2\right\}\\&\propto\exp\left\{-\frac{N}{2\sigma^2}(\bar{y}-\theta)^2\right\}\end{aligned}$$

其中，$\bar{y}=N^{-1}\sum_i y_i$，并且我们使用了 $\sum_i(y_i-\theta)^2=\sum_i(y_i-\bar{y}+\bar{y}-\theta)^2=\sum_i(y_i-\bar{y})^2+\sum_i(\bar{y}-\theta)^2$。乘法项不包含 θ，这被并入比例常值之中而被省略。频率学派

方法对数似然求关于 θ 的极大值,得出 MLE $\hat{\theta}=\bar{y}$。

此外,贝叶斯方法对 θ 的先验信息进行设定。从解析形式上看,一种方便的方式是选择正态先验,满足 $\theta\sim\mathcal{N}[\mu, \tau^2]$,其中对先验均值 μ 与先验方差 τ^2 的值进行设定。大的 τ^2 值表明,比其较小值具有更大的先验不确定性。于是,先验密度是:

$$\begin{aligned}\pi(\theta)&=(2\pi\tau^2)^{-1/2}\exp\{-(\theta-\mu)^2/2\tau^2\}\\&\propto\exp\{-(\theta-\mu)^2/2\tau^2\}\end{aligned}$$

其中,$(2\pi\tau^2)^{-1/2}$ 不含有 θ,它被并入比例因子之中。利用式(13.4),获得后验密度:

$$p(\theta|\mathbf{y})=\frac{\mathrm{L}(\mathbf{y}|\theta)\pi(\theta)}{\int_{-\infty}^{\infty}\mathrm{L}(\mathbf{y}|\theta)\pi(\theta)d\theta},\quad -\infty<\theta<\infty \tag{13.5}$$

分母确保了后验是正常的(也就是说,对它积分为 1)。就某些目的而言,可忽略分母,在此情况下,以 $p(\theta|\mathbf{y})\propto\mathrm{L}(\mathbf{y}|\theta)\pi(\theta)$ 进行研究。这时,对分子做如下扩展:

$$\begin{aligned}\mathrm{L}(\mathbf{y}|\theta)\pi(\theta)&=(2\pi\sigma^2)^{-N/2}\exp\left\{-\sum_{i=1}^{N}\frac{(y_i-\theta)^2}{2\sigma^2}\right\}(2\pi\tau^2)^{-1/2}\exp\left\{-\frac{(\theta-\mu)^2}{2\tau^2}\right\}\\&=(2\pi)^{-(N+1)/2}(\sigma^2)^{-N/2}(\tau^2)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\theta)^2-\frac{(\theta-\mu)^2}{2\tau^2}\right\}\end{aligned}$$

因为:

$$\sum_{i=1}^{N}(y_i-\theta)^2=\sum_{i=1}^{N}(y_i-\bar{y})^2+N(\bar{y}-\theta)^2$$

并注意到,式(13.5)的积分常值以及与 θ 独立的其他一些乘法常值都被并入比例常值之中,故有:

$$p(\theta|\mathbf{y})\propto\exp\left\{-\frac{N}{2\sigma^2}(\theta-\bar{y})\right\}\exp\left\{-\frac{1}{2}\frac{(\theta-\mu)}{\tau^2}\right\} \tag{13.6}$$

$$\propto\exp\left\{-\frac{1}{2}\left[\frac{(\theta-\mu)^2}{\tau^2}+\frac{(\bar{y}-\theta)^2}{N^{-1}\sigma^2}\right]\right\}$$

$$\propto\exp\left\{-\frac{1}{2}\left[\frac{(\theta-\mu_1)^2}{\tau_1^2}\right]\right\} \tag{13.7}$$

最后一行为 $\mathcal{N}[\mu_1, \tau_1^2]$ 分布的核,其中:

$$\mu_1=\tau_1^2(N\bar{y}/\sigma^2+\mu/\tau^2) \tag{13.8}$$

$$\tau_1^2=(N/\sigma^2+1/\tau^2)^{-1}$$

式(13.7)最后一行可通过完成平方而获得,若利用任意纯量 z, y, a_1, a_2, c_1 以及 c_2 的结果,有:

$$c_1(z-a_1)^2+c_2(z-a_2)^2=(c_1+c_2)\left(z-\left(\frac{c_1a_1+c_2a_2}{(c_1+c_2)^2}\right)\right)^2+\frac{c_1c_2}{(c_1+c_2)}(a_1-a_2)$$

其中,$z=\theta$, $a_1=\mu$, $a_2=\bar{y}$, $c_1=1/\tau^2$,而 $c_2=1/(N^{-1}\sigma^2+\tau^2)$。不含 θ 的项被省略。

总之,有下述内容:

数据：$y|\theta \sim \mathcal{N}[\theta, \sigma^2]$，$\sigma^2$已知。

先验：$\theta \sim \mathcal{N}[\mu, \tau^2]$，$\mu$，$\tau^2$设定。

后验：$\theta|\mathbf{y} \sim \mathcal{N}[\mu_1, \tau_1^2]$，$\mu_1$，$\tau_1^2$由式(13.8)给出。

后验均值(**posterior mean**) μ_1 是含有反映似然精度 σ^2/N 及先验 τ^2的先验均值 μ 与样本均值 $\bar{y}$ 的加权之和。贝叶斯的通常做法是，利用精度参数(**precision parameter**)概括可变性，而精度参数被定义为方差的倒数。这里的后验精度(**posterior precision**) τ_1^{-2}表示 $\bar{y}$ 的样本精度 N/σ^2与先验精度(**prior precision**) $1/\tau^2$之和，因此，精度可通过混合样本与先验信息而增大。

如果先验信息是不精确的，因而 $1/\tau^2$很小，分配给先验均值的权重相对于样本信息来说也就很小，从而先验在生成后验时起很小作用。类似地，当样本量增大时，样本信息同样占有优势，进而 N/σ^2相对于 $1/\tau^2$来说就大。后验分布趋于人们熟悉的渐近正态，只是贝叶斯结果是 $\theta \stackrel{a}{\sim} \mathcal{N}[\bar{y}, \sigma^2/N]$ ，而不是 $\bar{y} \stackrel{a}{\sim} \mathcal{N}[\theta, \sigma^2/N]$。

举一个具体例子，假定 $\sigma^2=100$，先验令 $\mu=5$ 且 $\tau^2=3$，而且容量 $N=50$ 的样本具有样本均值 $\bar{y}=10$。于是，似然是 $\mathcal{N}[10, 2]$，先验是 $\mathcal{N}[5, 3]$，由式(13.7)与式(13.8)知，后验为 $\mathcal{N}[8, 1.2]$。这些密度已画在图 13.1 中。后验均值位于先验均值与样本均值之间，而后验的方差既比先验方差小，又比似然方差小。

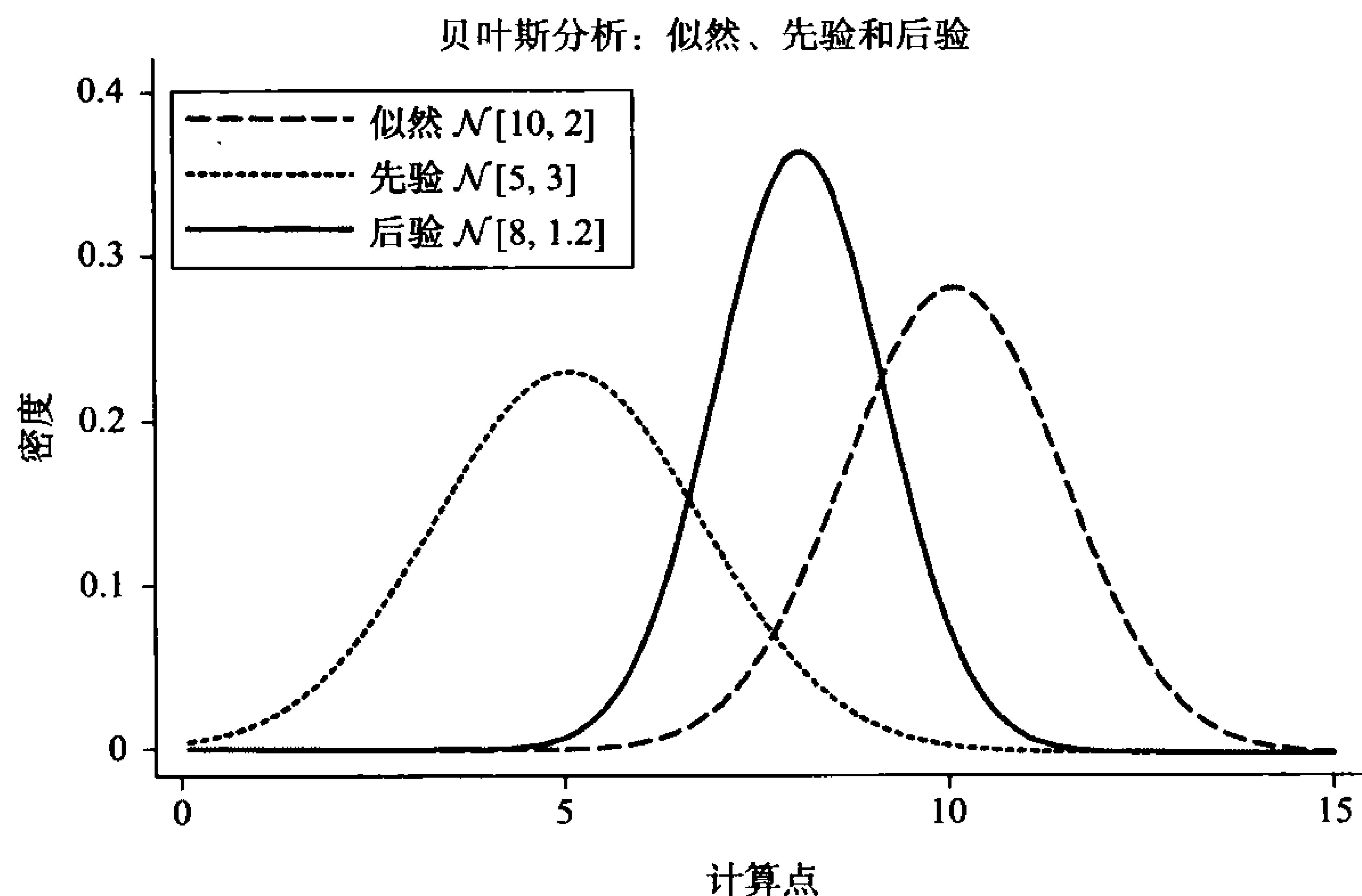


图 13.1 正态密度的均值参数的贝叶斯分析：正态似然(右边)、正态先验密度(左边)，以及所得到的后验密度(中间)。

13.2.3 *贝叶斯方法和非贝叶斯方法比较*

在频率学派方法与贝叶斯方法之间，找出其异同点是有益的。在参数频率学派方法的系统阐述中，似然函数是统计推断的一个重要基石。在合适正则条件下，MLE 是一致的且渐近正态的。估计量的抽样理论提供了有关估计数量或者其函数或者条件预测的概率表述的基础。关于参数的先验则被并入约束的 ML 估计中。

在贝叶斯分析中，数据生成过程及数据与参数的先验分布结合在一起，如表13.1所示。对这种先验分布的设定，将在13.2.4节加以详细讨论。在对现有数据进行分析，并建立在"已接收信息"基础之前，先验被嵌入到可能设定的信息里。利用贝叶斯定理，将先验信息与数据结合起来。

表13.1 贝叶斯分析:基本成分

成 分	公 式
抽样模型	出自 $f(y\|\boldsymbol{\theta})$ 的 iid $(y_1, \cdots, y_N)$
联合密度 / 似然	$f(y\|\boldsymbol{\theta})$, $L(\mathbf{y}\|\boldsymbol{\theta})$; $\boldsymbol{\theta} \in \boldsymbol{\Theta}$
先验分布	$\pi(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$
后验密度	$p(\boldsymbol{\theta}\|\mathbf{y}): \begin{cases} = f(\mathbf{y}\|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \Big/ \int f(\mathbf{y}\|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \\ \propto f(\mathbf{y}\|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\ \propto L(y\|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \end{cases}$
后验 pdf → 后验推断 →	参数估计 概率表述 预测 模型比较

运用这种方式得出了参数 $\boldsymbol{\theta}$ 的后验分布，可将其考虑成变换的似然函数。否则，给定数据，后验分布反映出我们的"先验修正"(**revised prior**)。当样本很小，而且或许相对没有什么信息价值，后验分布看起来好像是一个先验分布，但当样本很大时，后验分布将反映出数据的特性。

13.2.4 *先验设定*

贝叶斯分析需要对 dgp $f(\mathbf{y}|\boldsymbol{\theta})$与先验 $\pi(\boldsymbol{\theta})$进行设定。通常，dgp 被设定成与完全参数基于似然分析中所使用的相同。对于二值结果来说，设定 logit 或 probit 模型，对于计数数据来说，设定泊松模型或负二项式模式等。

与经典分析相比，由贝叶斯分析所引起的原则性挑战是，需要额外地对先验分布进行设定。其结果会随先验选择不同而变化，因为各种不同先验会导致不同的后验分布，除非样本量足够大到使得样本信息占有优势。

一种方法是选取先验分布，以使它对后验分布具有很小的影响，因此，其结果本质上是建立在样本数据的基础之上。一种可供选择的方法是，当保证具有很强的先验信息可利用时，就是去设定反映这种信息的先验。上述两种方法，尤其是后者，在历史上被后验分布处理性问题所束缚，可是，这一点因当今目前计算进步而不在考虑范围内。一种流行的中间方法是运用层次先验(**hierarchical prior**)，其关于参数不确定性可利用概率函数表述，而它们本身涉及其他参数也不能断定。

非信息先验

非信息先验(**noninformative prior**)是指那种对所得到的后验分布具有很小影响的先验。

获得非信息先验的一种明显方法是,使用一致先验(**uniform prior**),对于所有 $\boldsymbol{\theta}$ 满足 $\boldsymbol{\pi}(\boldsymbol{\theta})=c$,其中,$c>0$ 表示常值,因为这会对 $\boldsymbol{\theta}$ 的所有可能值设置相等权数。

一致先验的一个缺点是,若在参数 $\boldsymbol{\theta}$ 为有界的背景下使用,此先验就是一种非正常密度(**improper density**),因为必然有 $\int \pi(\boldsymbol{\theta})d\boldsymbol{\theta}=\infty$。于是,得到的后验分布也可能是非正常的,尽管在几个重要例子中,后验分布仍然是正常的。

一致先验的另一个缺点是,它对重参数化不是不变的。例如,对于纯量参数 $\theta>0$ 来说,一种可供选择的密度 y 的明显参数化是依照参数 $\gamma=\ln\theta$,进而 $-\infty<\gamma<\infty$。当 θ 服从一致先验,$\pi(\theta)=c$,γ 的对应先验 $\pi^*(\gamma)$ 就不是一致的,因为 $\pi^*(\gamma)=\pi(\theta)|d\theta/d\gamma|=ce^{\gamma}$。尽管对一个参数化来说好像没有什么信息,但对另一个参数化而言,该先验却是有信息价值的。

一致先验能通过设定具有非常大方差的正常先验而得以仿效。例如,假定纯量 θ 服从 $\mathcal{N}[\mu,\tau^2]$ 先验,其中,τ^2 是非常大的。从而,对于可能通过数据支撑的 θ 值来说,先验 $\pi(\theta)\simeq 1/(2\pi\tau^2)$,即一个常值,因为 $\exp[-(\theta-\mu)/2\tau^2]\simeq 1$。重要的是注意到,这种明显方法与一致先验具有相同的缺陷,称为非确定的(**vague**)或散开的(**diffuse**)或平坦的先验(**flat prior**)。对重参数化而言,该方法不是不变的。

不过,一种广泛使用的非信息先验是杰弗里斯先验(**Jeffreys' prior**):

$$\pi(\boldsymbol{\theta})\propto|\mathcal{I}(\boldsymbol{\theta})|^{1/2} \tag{13.9}$$

其中,对于向量 $\boldsymbol{\theta}$ 来说,$|\mathcal{I}(\boldsymbol{\theta})|$ 表示信息矩阵 $\mathcal{I}(\boldsymbol{\theta})=-\mathrm{E}[\partial^2\mathcal{L}/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}']$ 的行列式,满足 $\mathcal{L}=\ln \mathrm{L}(\mathbf{y}|\boldsymbol{\theta})$。杰弗里斯先验是以先驱贝叶斯·哈罗德·杰弗里斯(Bayesian Harold Jeffreys)命名的,它对重参数化或模型参数的变换来说,具有不变性(**invariance**),因此,不管选取的特殊参数化如何,都将得出一样的先验信息。

为检证杰弗里斯规则,为了简单起见,我们考察纯量参数情况。已知变换 $\gamma=h(\theta)$,$\partial\mathcal{L}/\partial\gamma=\partial\mathcal{L}/\partial\theta\times\partial\theta/\partial\gamma$,并且:

$$\frac{\partial^2\mathcal{L}}{\partial\gamma^2}=\frac{\partial^2\mathcal{L}}{\partial\theta^2}\left(\frac{\partial\theta}{\partial\gamma}\right)^2+\frac{\partial\mathcal{L}}{\partial\theta}\frac{\partial^2\theta}{\partial\gamma^2}$$

对样本密度取期望,同时注意到,$\mathrm{E}[\partial\mathcal{L}/\partial\theta]=0$,由似然得分性质可得:

$$\mathcal{I}(\gamma)=\mathcal{I}(\theta)\left(\frac{\partial\theta}{\partial\gamma}\right)^2$$

由此可得:

$$|(\mathcal{I}(\gamma)|^{1/2}=|\mathcal{I}(\theta)|^{1/2}\left|\frac{\partial\theta}{\partial\gamma}\right|$$

通常,θ 的先验 $\pi(\theta)$ 蕴含,关于 γ 的先验为 $\pi^*(\gamma)=\pi(\theta)\times|d\theta/d\gamma|$。若专门研究先验(13.9),则得出 $\pi^*(\gamma)\propto|\mathcal{I}(\theta)|^{1/2}\times|d\theta/d\gamma|$,但这正是人们所期望的 $|\mathcal{I}(\gamma)|^{1/2}$。

举一个例子,假定 $y\sim\mathcal{N}[\mu,\sigma^2]$,并考察三种情况。第一种情况,当 μ 是未知参数且 σ^2 是已知的,关于 μ 的信息测度是 $\mathcal{I}(\mu)=N/\sigma^2$,而杰弗里斯先验 $|\mathcal{I}(\mu)|^{1/2}\propto c$ 为一个常值,因为,这里 σ^2 是已知的。注意,该先验是非正常先验。第二种情况,当 σ^2 是未知的且 μ 是已知的,关于 σ^2 的信息测度是 $\mathcal{I}(\sigma^2)=N/(2\sigma^4)$,而杰弗里斯先

验为 $|\mathcal{I}(\sigma^2)|^{1/2}\propto\sigma^{-2}$。第三种情况，当 μ 是未知的，σ^2 也是未知的，信息矩阵 $|\mathcal{I}(\mu,\sigma^2)|=(N/\sigma^2)(N/2\sigma^4)=N^2/2\sigma^6$。因此，杰弗里斯规则蕴含联合先验 $\pi(\mu,\sigma^2)\propto\sigma^{-3}$。注意，这不同于将杰弗里斯规则应用到 μ 与 σ^2 的各自先验上所得到的结果，因为 $\pi(\mu)\propto c$ 与 $\pi(\sigma^2)\propto\sigma^{-2}$，得出 $\pi(\mu)\pi(\sigma^2)\propto\sigma^{-2}$。

当没有明显的备选先验可利用时，杰弗里斯规则能作为生成先验的一种方法。可是，文献似乎没有解决规则是否产生非信息先验的问题，而且如果有的话，又有什么意义？更进一步，由前面的例子，杰弗里斯先验可能是非正常的，这一点很明显，从而得出非正常后验。

共轭先验

当设定正常先验时，或是作为信息先验或是作为散开先验，已知数据的设定样本密度，选择会产生关于后验的容易处理的"良好"解析表达式，诸如式(13.7)的函数形式是方便的。若样本与先验密度出自自然共轭对(**natural conjugate pair**)，即定义成具有下述性质：样本密度与先验分布及后验分布全部处于相同类型的密度中。那么，这类容易处理的结果大多经常会出现，于是，先验被称为自然共轭先验(**natural conjugate prior**)。13.2.2 节已给出一个例子，对于正态分布数据来说，其均值的正态先验导致后验分布也是正态的。

指数族基本上是具有自然共轭对的唯一密度类型。指数族的一个参数成员具有下述密度；单个观测值表述成：

$$f(y|\theta)=\exp\{a(\theta)+b(y)+c(\theta)u(y)\} \qquad (13.10)$$
$$\propto\exp\{a(\theta)+c(\theta)u(y)\}$$

其中，各不相同的函数 $a(\cdot)$、$c(\cdot)$ 以及 $u(\cdot)$ 会产生族中不同密度，而 $b(\cdot)$ 表示正规化常值。例如，设 $c(\theta)=\mu/\sigma^2$、$a(\theta)=-\mu^2/2\sigma^2$ 以及 $u(y)=y$，则得到 $\mathcal{N}[\mu,\sigma^2]$ 分布的核(关于 σ^2 是已知)。注意到，设 $u(y)=y$ 会产生线性指数族，5.7.3 节曾经以某种详细方式阐述过。更一般地讲，若 $\boldsymbol{\theta}$ 表示一个向量，则 $c(\theta)u(y)$ 可用 $\mathbf{c}(\boldsymbol{\theta})'\mathbf{u}(y)$ 代替，其中，$\mathbf{u}(\cdot)$ 通常具有与 $\boldsymbol{\theta}$ 一样的维数。

对于容量为 N 的随机样本来说，由指数族，得出样本密度：

$$\mathrm{L}(\mathbf{y}|\theta)\propto\exp\{Na(\theta)+c(\theta)t(\mathbf{y})\} \qquad (13.11)$$

其中，$t(\mathbf{y})=\sum_i u(y_i)$。考察下述关于 θ 的先验：

$$\pi(\theta|\beta,\alpha)\propto\exp\{\beta a(\theta)+\alpha c(\theta)\} \qquad (13.12)$$

其中，α 与 β 均是先验的设定参数，而函数 $a(\cdot)$ 与 $c(\cdot)$ 均与式(13.10)的那些一样。当将 α 看成固定的，这个密度就是 θ 的指数族密度。应用贝叶斯定理，经过简化得到：

$$p(\theta|\mathbf{y})\propto\mathrm{L}(\mathbf{y}|\theta)\pi(\theta|\beta,\alpha) \qquad (13.13)$$
$$\propto\exp\{(\beta+N)a(\theta)+(\alpha+t(\mathbf{y}))c(\theta)\}$$

容易验证，这与式(13.12)的最初先验具有一样的核。一旦将后验与样本密度相比，则显示先验可被看成提供一个额外 β 观测值 $\mathbf{y}_p$，比如说，满足 $t(\mathbf{y}_p)=\alpha$。

表 13.2 阐述某些标准的共轭族,其有关的密度已在附录 B 中提供。伽玛包括指数与卡方作为其特殊情况。同理,负二项式的、一致的以及帕累托的似然都具有共轭先验密度。

表 13.2　共轭族:重要例子

分布	样本密度	共轭先验密度
正态分布	$\mathcal{N}[\theta, \sigma^2]$	$\theta \sim \mathcal{N}[\mu, \tau^2]$
正态分布	$\mathcal{N}[\mu, 1/\theta^2]$	$\theta \sim \mathcal{G}[\alpha, \beta]$
二项分布	$\mathcal{B}[N, \theta]$	$\theta \sim \mathrm{Beta}[\alpha, \beta]$
泊松分布	$\mathcal{P}[\theta]$	$\theta \sim \mathcal{G}[\alpha, \beta]$
伽玛分布	$\mathcal{G}[v, \theta]$	$\theta \sim \mathcal{G}[\alpha, \beta]$
多项式分布	$\mathcal{MN}[\theta_1, \cdots, \theta_k]$	$\theta_1, \cdots, \theta_k \sim \mathrm{Dirichlet}[\alpha_1, \cdots, \alpha_k]$

共轭先验的一个引人注目之处是,得到的计算结果在计算形式与解析形式上均具有简单性。不过,运用共轭先验是受到限制的,而且对典型研究者来说,当利用的资源受到相当限制时,对是否正确地利用它做出判断,现在比过去更缺少强有力的依据。

拥有与先验同一类型的后验的另一个优点是,后验很容易用先验来代替,并作为后面分析的一个(基于数据)新先验。若将先验解释成"已接受信息",则人们从研究中得到的后验作为下面探索的先验。

分层先验

当先验的参数自身被建模成一个分布时,就产生分层先验(**hierarchical priors**)。出现在这类"关于先验的先验"当中的参数,被称为超参数(**hyperparameters**)。

如同 13.2.1 节一样,数据具有联合密度 $\mathrm{L}(\mathbf{y}|\boldsymbol{\theta})$,但现在 $\boldsymbol{\theta}$ 的先验依赖于参数 $\boldsymbol{\tau}$,比如说,$\boldsymbol{\tau}$ 是随机的而不是固定的。因而,$\boldsymbol{\theta}$ 的先验是 $\pi(\boldsymbol{\theta}|\boldsymbol{\tau})$,其中,参数 $\boldsymbol{\tau}$ 同样具有一个先验 $\pi(\boldsymbol{\tau})$。联合先验是 $\pi(\boldsymbol{\theta},\boldsymbol{\tau})=\pi(\boldsymbol{\theta}|\boldsymbol{\tau})\pi(\boldsymbol{\tau})$,由贝叶斯规则可得,其联合后验:

$$p(\boldsymbol{\theta},\boldsymbol{\tau}|\mathbf{y}) \propto \mathrm{L}(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\tau})\pi(\boldsymbol{\tau})$$

关注内容通常是 $\boldsymbol{\theta}$ 的边缘后验,这通过联合后验对 $\boldsymbol{\tau}$ 进行积分而得到。先验 $\pi(\boldsymbol{\tau})$ 的设定参数称为超参数。作为一种可供选择的方式,这种参数同样可以是已知先验,在此情况下,引入另一种分层,得出联合先验 $\pi(\boldsymbol{\theta}|\boldsymbol{\tau})\pi(\boldsymbol{\tau}|\phi)\pi(\phi)$ 等。最近,贝叶斯分析计算方法进步,特别是吉布斯抽样器,都很好地适用于分层先验,因为这些方法有递推结构。

可将分层先验看成是,在典型设置背景下随机系数模型的贝叶斯类似形式。例如,对于 iid 计数数据来说,假定 $y_i \sim \mathcal{P}[\theta_i]$,其中,泊松参数现在是随机的。关于 θ_i 的一个方便分布是共轭伽玛分布,因此 $\theta_i \sim \mathcal{G}[\alpha, \beta]$。一种典型的方法是,通过极大似然法估计 α 与 β。非分层贝叶斯模型对 α 与 β 加以设定,并获得关于 θ_i 的后验。分层贝叶斯模型则对 α 与 β 设定其先验,例如作为共轭形式的伽玛,并在求关于 θ_i 的边际后验之前,首先要求关于 θ_i、α 和 β 的联合后验。

在分层模型(**hierarchical models**)的背景下,自然出现分层先验,这类模型统称为多层模型(**multilevel models**)。这种模型广泛用于经典设置背景下利用特定目的软件情况。在贝叶斯设置背景下,林德利和史密斯(Lindley and Smith, 1972)对分层回归模型分析做出了早期的研究工作。只要被分析的数据自然归入层、组或小层(**layers**)时,分层模型自然有吸引力,而且人们希望看到,关注关系在分组参数(**groupwise parameter**)上的变异。例如,测验分数的观测值可来自特定年级与学校的学生。对测验分数的建模能包含随不同个体变化而定义的个体特征、随不同年级而变化的班级特征,以及仅随不同学校而变化的学校特征。由于这类数据将会涉及观测值的集群,所以这个专题也将在第 24 章讨论。这类模型与面板数据的随机效应具有密切关系。

举一个例子,假定数据可归入 J 个组,同时 y 的总体均值会随不同组而变化。对于第 j 组的个体 i 来说,假定 $y_{ij}\sim\mathcal{N}[\theta_j,\sigma^2]$,其中,为了简单起见,假定 σ^2 是已知的。从而,第 j 个组中的样本均值 $\bar{y}_{ij}\sim\mathcal{N}[\theta_j,\sigma^2/N_j]$,其中,$N_j$ 表示组的个体数目,并假设独立性成立。例如,分层模型设定均值 θ_j 具有先验 $\theta_j\sim\mathcal{N}[\mu,\tau^2]$,其中,对于较高层先验的参数 μ 与 τ^2 来说,要设定另外的先验。

敏感性分析

在频率学派分析中,人们会考虑用于系统建立估计模型的一系列准确的先验约束。例如,在一个或多个约束集合下,对模型进行估计,而其结果与来自对先验假设实施估计的敏感性思想相比较。

同样的逻辑及方法,可运用于贝叶斯分析。人们不必使先验严格正确,而人们能实施敏感性分析,研究后验是如何随先验的不同选取而变化的。类似地,人们能改变关于数据生成过程的假设,并分析后验信息会怎样响应变化。

13.2.5 与后验有关的密度和测量

贝叶斯分析建立在后验分布的基础上。为了方便起见,贝叶斯回归结果通常只报告概括性测量,诸如后验矩、分位数或 $\boldsymbol{\theta}$ 分量的边缘分布。然而,后验分布也可用于预测与概率表述,对此本节将详细阐述;它还可用于模型比较,这将在 13.8 节加以阐述。

边缘后验

通常,$\boldsymbol{\theta}$ 是多维数的,用 $\boldsymbol{\theta}'=(\theta_1,\cdots,\theta_q)$ 表示,人们关注的内容可以是 $\boldsymbol{\theta}$ 的个体成分后验分布。第 k 个参数 θ_k 的边缘后验密度(**marginal posterior density**),通过对 $\boldsymbol{\theta}$ 的联合后验中剩余 $(q-1)$ 个全部分量进行积分而获得。正式地讲,这表示成 $p(\theta_k|\mathbf{y})$,并通过计算 $(q-1)$ 重积分得到:

$$
\begin{aligned}
p(\theta_k|\mathbf{y}) &= \int p(\theta_1,\cdots,\theta_p|\mathbf{y})d\theta_1\cdots d\theta_{k-1}d\theta_{k+1}\cdots d\theta_q \qquad (13.14)\\
&= \int p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}_{-k}
\end{aligned}
$$

其中,第二行中更简洁的记号包含 $\boldsymbol{\theta}_{-k}$,$\boldsymbol{\theta}_{-k}$ 表示 $\boldsymbol{\theta}$ 去掉 θ_k 之后的所有元素。通常,边缘后验密度是非对称的且不必是单峰的。特别地,当边缘后验密度远远违背对

称单峰分布时，画出后验图形是有用的。

后验矩

经典回归输出会报告参数估计值与标准误差。对于贝叶斯回归来说，人们会类似报告每个参数的边缘后验密度的均值或中位数、标准差。

点估计

在经典分析里，存在未知真实参数值 $\boldsymbol{\theta}_0$，使数据生成过程是 $f(\mathbf{y}|\boldsymbol{\theta}_0)$，并求其点估计，它是 $\boldsymbol{\theta}_0$的一个良好估计。与之相比，贝叶斯分析关注内容是 $\boldsymbol{\theta}$ 的整个分布，它既由 $\boldsymbol{\theta}_0$决定，又由关于 $\boldsymbol{\theta}_0$的先验信念决定。

因此，贝叶斯分析很少强调点估计。不过，为了方便起见，后验均值与后验中位数被广泛报告出来作为点估计。通过设定损失函数，获得参数的最优点估计；参见 13.2.7 节。

后验区间

一旦获得后验分布，它可用于做出类似于频率学派分析的概率表述。特别地，我们考察贝叶斯置信区间与区域。

对于第 k 个参数来说，$100(1-\alpha)\%$后验密度区间 $\mathcal{R}(\theta_k)$（**posterior density interval**）是 θ_k 以后验概率 α 落入的任何一个区间，或正式地：

$$1-\alpha=\Pr[\theta_k\in\mathcal{R}(\theta_k)|\mathbf{y}]=\int_{\mathcal{R}(\theta_k)}p(\theta_k|\mathbf{y})d\theta \tag{13.15}$$

对应于这个概率，存在许多区域。一个最简单的后验区间是，位于 $\alpha/2$ 与$(1-\alpha/2)$分位数之间的区间，比如在 2.5 分位数与 97.5 分位数之间。一个更复杂的情况是最高后验密度区间[**highest posterior density (HPD) interval**]，它要满足式(13.15)以及下述另外条件：$\mathcal{R}(\theta)$中没有任何一点比其区域外任何点的概率密度小。当后验是多峰时，这一区间不必是连接的，同时它不同于较简单的区间，除非后验是对称的且单峰的。

可以将这些区间推广到区域上。一个 $100(1-\alpha)\%$的最高后验密度区域 $\mathcal{R}(\theta)$，使得：

$$1-\alpha=\Pr[\boldsymbol{\theta}\in\mathcal{R}(\theta)|\mathbf{y}]=\int_{\mathcal{R}(\theta_k)}p(\boldsymbol{\theta}|\mathbf{y})d\theta \tag{13.16}$$

贝叶斯方法的引人注目之处是，与频率学派分析置信区间相比，后验区间解释起来更加简单。当 θ_k 的 95%后验区间是(1, 4)时，则 θ_k 以后验概率 0.95 位于 1 与 4 之间。与之相比，对于频率学家来说，θ_k 的 95%置信区间等于(1, 4)，我们只能说，如果可能的话，以许多各种不同样本进行重复分析，会得到一些不一样的置信区间，那么这些置信区间的 95%将包括 θ_k的真实值。

假设检验

在贝叶斯背景下，假设检验很少受到注意。如同在对点估计的讨论中所提及的，人们关注的内容不是去确定真实参数值 $\boldsymbol{\theta}_0$。相反，关注的内容是，已知数据与先验时 $\boldsymbol{\theta}$ 可能取值的范围分布。对于模型比较来说，参见 13.8 节。

条件后验密度

已知 θ_j 时，θ_k 的条件后验密度(**conditional posterior density**)由联合后验密度

与边缘后验密度来获得，即：

$$p(\theta_k|\theta_j, \theta_j \in \boldsymbol{\theta}_{-k}, \mathbf{y}) = \frac{p(\theta_k, \theta_j|\mathbf{y})}{p(\theta_j|\mathbf{y})} \tag{13.17}$$

特别关注内容及具有重要意义的是，q 个条件分布集合 $p(\theta_k|\boldsymbol{\theta}_{-k}), k=1,\cdots,q$，这也是众所周知的完全条件分布（**full conditional distributions**）。对于后面几节将阐述的联合后验分布来说，这些在现代计算方法上起着重要作用。

式(13.15)与式(13.17)所定义的边缘后验与条件后验，可从个体参数推广到分块参数上（**blocks of parameters**）。

边缘似然

边缘似然（**marginal likelihood**）的边缘概率是贝叶斯法则的分母，并被定义成：

$$f(\mathbf{y}) = \int \mathrm{L}(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{13.18}$$

它是似然的期望值，即 $\mathrm{E}[\mathrm{L}(\mathbf{y}|\theta)]$，这里的期望是关于先验密度的。边缘似然构成贝叶斯推断的基础（参见 13.8 节），因为它包含关于数据支持先验的信息。

后验预测密度

考察单个观测值 y^p 的样本外预测。这具有密度 $f(y^p|\boldsymbol{\theta})$，其中，$\boldsymbol{\theta}$ 表示未知的。y^p 的后验预测密度（**posterior predictive density**），通过 $\boldsymbol{\theta}$ 的后验概率分布对该密度加权，得到：

$$f^p(y^p) = \int f(y^p|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \tag{13.19}$$

正如回归模型一样，协方差出现在似然函数中时，这些密度也同样以它们为条件。

13.2.6 *后验大样本特性*

如同 13.2.2 节例子所阐明的，当样本变大时，甚至有信息的先验对后验的影响会消失。这是下面陈述的根基：渐近似然支配着推断，或先验分配的权数本质上会随样本量增大而趋于 0。

由于认识到，可运用后验分布，对后验的渐近近似是人们所关注的，因为它能用于代替真实的有限样本后验分布。由于渐近后验等于似然，所以这种近似很容易获得。我们沿着格尔曼等人（Gelman et al.，1995）的线索展开，对于更详细的内容，请读者参考他们的书。

为了简单起见，假定观测值是 iid 的。于是，后验对数为：

$$\sum_{i=1}^{N} \ln p(\boldsymbol{\theta}|y_i) = \ln \pi(\boldsymbol{\theta}) + \sum_{i=1}^{N} \ln f(y_i \mid \boldsymbol{\theta}) \tag{13.20}$$

该表达式清楚表明，在大样本中，后验是由其似然贡献所控制，因为先验对后验的贡献保持固定，而样本对后验的贡献却随 N 而增大。

假定后验 $p(\boldsymbol{\theta}|\mathbf{y})$ 是单峰的且渐近对称的。考察后验众数的渐近性质，用 $\hat{\boldsymbol{\theta}}$ 表示，于是，它是后验的局部与全局最大值。

为了建立 $\hat{\boldsymbol{\theta}}$ 的一致性，注意到，当 $N\to\infty$时，后验形式收敛到 MLE，因为式(13.20)的第二项占据控制。因此，若 MLE 是一致的，则后验众数是一致的。所以，如果关于 y 的数据生成过程具有密度 $f(y|\boldsymbol{\theta}_0)$，同时关于 ML 估计的通常正则条件得以满足，那么 $\hat{\boldsymbol{\theta}}\xrightarrow{p}\boldsymbol{\theta}_0$。

为了获得 $\hat{\boldsymbol{\theta}}$ 的渐近分布，考察后验对数密度在后验众数附近的二阶泰勒级数序列展开式。从而当在后验众数处进行计算时，因$\partial\ln p(\boldsymbol{\theta}|\mathbf{y})/\partial\boldsymbol{\theta}=\mathbf{0}$，故可以简化为：

$$\ln p(\boldsymbol{\theta}|\mathbf{y})\simeq\ln p(\hat{\boldsymbol{\theta}}|\mathbf{y})+\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})'\left[\left.\frac{\partial^2\ln p(\boldsymbol{\theta}|\mathbf{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\theta=\hat{\theta}}\right](\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) \quad (13.21)$$

并假定 $\boldsymbol{\theta}$ 的第三阶导数与更高阶导数能被渐近忽略。定义：

$$\mathcal{I}(\hat{\boldsymbol{\theta}})=-\left.\frac{\partial^2\ln p(\boldsymbol{\theta}|\mathbf{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\theta=\hat{\theta}}$$

是建立后验密度 $\ln p(\boldsymbol{\theta}|\mathbf{y})$ 基础上的可观测信息，在后验众数处的计算值。于是，对式(13.21)进行指数化，得到：

$$p(\boldsymbol{\theta}|\mathbf{y})\propto\exp\left(-\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})'\mathcal{I}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\right)$$

这是多元变量正态分布的核，其均值为 $\hat{\boldsymbol{\theta}}$ 且方差矩阵为 $\mathcal{I}(\hat{\boldsymbol{\theta}})^{-1}$。由此可得，后验：

$$\boldsymbol{\theta}|\mathbf{y}\overset{a}{\sim}\mathcal{N}[\hat{\boldsymbol{\theta}},\mathcal{I}(\hat{\boldsymbol{\theta}})^{-1}] \quad (13.22)$$

当样本量 N 增大时，后验的似然成分占据控制地位，而先验影响却变得可以忽略。在这种情况下，可用 MLE 代替众数 $\hat{\boldsymbol{\theta}}$，作为似然密度的众数。从而，得到有时被称为贝叶斯中心极限定理(**Bayesian central limit theorem**)的结果[盖默曼(Gamerman, 1997)]。从渐近形式上看，频率学派推断与贝叶斯推断都将建立在同样多元变量的正态分布的基础上，因此它们之间的不一致不应是显著的。

文献中将这个结果称为伯恩斯坦—冯·米泽斯定理(**Bernstein - von Mises Theorem**)；特雷恩(Train, 2003, 第 12 章)对该定理的三个成分提供了通俗易懂的讨论。这些成分包括：(1)后验均值依概率收敛到极大似然估计量的结果；(2)它具有极限正态分布；(3)后验均值的极限分布与极大似然估计量的极限分布是一样的。在贝叶斯中心极限定理中，这些结果全部是含蓄的(不言明的)，对于那些想要在估计与推断中应用似然原理的人，该定理与他们息息相关，应密切关注。

前面的讨论会蕴含贝叶斯与基于似然方法在本质上产生相似的结果吗？对两种方法的选取大部分可能会是一个计算效率问题吗？然而，文献中存在一系列论文，不仅证明这两种方法得出相似结果，而且证明贝叶斯方法在计算上常常更有效。

13.2.7 *贝叶斯决策分析*

已知完全后验分布 $p(\boldsymbol{\theta}|\mathbf{y})$，应报告 $\boldsymbol{\theta}$ 的哪个点估计呢？这个问题已在 4.2 节研究过，例如，对于 y 的最佳预测来说，利用平方误差损失。相反，这里考察 $\boldsymbol{\theta}$ 的最佳估计，例如，利用二次损失。

设 $L(\boldsymbol{\theta},\hat{\boldsymbol{\theta}})$ 表示设定的损失函数，其中，$\hat{\boldsymbol{\theta}}$ 表示未知 $\boldsymbol{\theta}$ 的估计值。损失是未知的，因为它依赖于 $\boldsymbol{\theta}$，而 $\boldsymbol{\theta}$ 是未知的。不过，我们能求损失关于 $\boldsymbol{\theta}$ 的期望值，与经典分析不同，这是因为贝叶斯分析提供 $\boldsymbol{\theta}$ 的分布。最优估计量（**optimal estimator**）$\hat{\boldsymbol{\theta}}_{OPT}$ 是求期望后验损失极小化（**minimizes expected posterior loss**）的估计量 $\hat{\boldsymbol{\theta}}$，或者：

$$\min_{\hat{\theta}} E[L(\boldsymbol{\theta},\hat{\boldsymbol{\theta}})] = \min_{\hat{\theta}} \int L(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \tag{13.23}$$

与各种不同的 $(\boldsymbol{\theta},\hat{\boldsymbol{\theta}})$ 相联系的损失是，通过后验概率 $p(\boldsymbol{\theta}|\mathbf{y})$ 进行加权。

可以证明，后验均值是在二次损失 $L(\boldsymbol{\theta},\hat{\boldsymbol{\theta}})=(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})'(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})$ 下的最优估计量。可是，若使用绝对误差，即 $L(\boldsymbol{\theta},\hat{\boldsymbol{\theta}})=|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}|$，后验中位数则是最优估计量。一旦建立起后验分布，这些点估计或以解析形式计算，或以数值形式计算。

在某些条件下，可以证明，求期望后验损失等价于求期望后验风险（**expected posterior risk**）极小化。风险函数来自总体 $\mathbf{y}$ 的假设样本对可能损失求平均，所以：

$$R(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}) = \int L(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}) f(\mathbf{y}|\boldsymbol{\theta}) d\mathbf{y}$$

为避免损失函数与似然函数之间相混淆，本节与下一节方程式中，均使用 $f(\mathbf{y}|\boldsymbol{\theta})$ 等价于似然函数 $L(\mathbf{y}|\boldsymbol{\theta})$。期望后验分布在参数 $\boldsymbol{\theta}\in\boldsymbol{\Theta}$ 的各种不同值上进行平均，而参数 $\boldsymbol{\theta}$ 可借助于后验密度进行加权，所以：

$$\begin{aligned} E[R(\boldsymbol{\theta},\hat{\boldsymbol{\theta}})] &= \int_{\boldsymbol{\Theta}} \left\{ \int L(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}) f(\mathbf{y}|\boldsymbol{\theta}) d\mathbf{y} \right\} p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \\ &= \int \left\{ \int_{\boldsymbol{\Theta}} L(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \right\} f(\mathbf{y}|\boldsymbol{\theta}) d\mathbf{y} \\ &= \int E[L(\boldsymbol{\theta},\hat{\boldsymbol{\theta}})] f(\mathbf{y}|\boldsymbol{\theta}) d\mathbf{y} \end{aligned} \tag{13.24}$$

其中，第一个等式的外面积分是针对 $\boldsymbol{\theta}$ 区域进行，第二个等式中的积分次序是可交换的，而第三个等式则为结论。这些运算均假定，$L(\boldsymbol{\theta},\hat{\boldsymbol{\theta}})$ 与 $p(\boldsymbol{\theta}|\mathbf{y})$ 上的一些约束都得以满足。例如，$p(\boldsymbol{\theta}|\mathbf{y})$ 必是正常密度函数，而损失函数必是可积的。因此，期望风险将是有界的，并对它求极小值是一种定义良好的运算。

前面的讨论建立了著名且重要的结果，即贝叶斯估计量在使其对设定损失函数的期望风险求极小值的意义上是合理的。

13.3 线性回归贝叶斯分析

由于线性回归分析是一个熟悉的专题，它为进入更一般非线性模型的研究提供了有益的途径。假定数据由标准线性回归模型

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{u}$$

生成，其中，$\mathbf{X}$ 表示弱外生的回归元 $N\times K$ 列满秩的矩阵。假定误差是独立的、同方差的且服从正态分布，满足 $\mathbf{u}\sim\mathcal{N}[\mathbf{0},\sigma^2\mathbf{I}_N]$。因此，样本条件密度是 $\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\sigma^2\sim\mathcal{N}[\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I}_N]$。我们的解释遵循泽尔纳（Zellner，1971）的线索。

我们依次研究非信息先验与信息先验。在这两种情况下，经过某种相当多的代数运算，能够获得关于后验的闭型表达式。对于非信息先验来说，将会看到 OLS 估计量作为后验分布的均值，具有贝叶斯解释。在信息先验情况下将会看到，后验矩是样本均值与先验均值的加权函数。

后面几节阐述比较容易处理模型的方法，尽管如此，分析仍可简化，如果结果类似于这一节给出的那些结果，它们能应用于模型的某些子成分。

13.3.1 *非信息先验*

对于非信息先验来说，我们使用杰弗里斯先验。由 13.2.4 节知，对于 $y \sim \mathcal{N}[\mu,\sigma^2]$来说，关于 μ 的这个先验（给定 σ^2为已知）是一个常值，而关于 σ^2的先验（给定 μ 为已知）与 σ^2成比例。就回归情况而言，这可推广到关于 β_j 的常值先验上，$j=1,\cdots,K$，因此 $\pi(\beta_j)\propto c$，而且关于 σ^2的先验是 $\pi(\sigma^2)\propto 1/\sigma^2$。先验将 β_j 的所有值看成相等的，而将 σ^2的较小值看成更大一些。若假定 β 与 σ^2的独立性，则联合先验是：

$$\pi(\boldsymbol{\beta},\sigma^2)\propto 1/\sigma^2$$

似然函数能重新写成：

$$\begin{aligned}
\mathrm{L}(\boldsymbol{\beta},\sigma^2|\mathbf{y},\mathbf{X}) &= (2\pi\sigma^2)^{-N/2}\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\} \qquad (13.25)\\
&\propto(\sigma^2)^{-N/2}\exp\left(-\frac{1}{2\sigma^2}\{\hat{\mathbf{u}}'\hat{\mathbf{u}}+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\}\right)\\
&\propto(\sigma^2)^{-N/2}\exp\left(-\frac{1}{2\sigma^2}(N-K)s^2+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right.
\end{aligned}$$

其中，$\hat{\boldsymbol{\beta}}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$，而 $\hat{\mathbf{u}}=\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}$；第二行运用 $\mathbf{y}-\mathbf{X}\boldsymbol{\beta}=\hat{\mathbf{u}}-\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})$ 且 $\mathbf{X}'\hat{\mathbf{u}}=\mathbf{0}$；第三行运用 $s^2=\hat{\mathbf{u}}'\hat{\mathbf{u}}/(N-K)$。

将式(13.15)的似然与先验结合起来，得出其后验密度：

$$\begin{aligned}
&p(\boldsymbol{\beta},\sigma^2|\mathbf{y},\mathbf{X}) \qquad (13.26)\\
&\propto\left(\frac{1}{\sigma^2}\right)^{N/2}\exp\left(-\frac{1}{2\sigma^2}\{(N-K)s^2+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\}\right)\frac{1}{\sigma^2}\\
&\propto\left(\frac{1}{\sigma^2}\right)^{N/2+1}\exp\left(-\frac{1}{2\sigma^2}\{(N-K)s^2+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\}\right)\\
&\propto\left\{\left(\frac{1}{\sigma^2}\right)^{K/2}\exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'(\sigma^2(\mathbf{X}'\mathbf{X})^{-1})^{-1}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right)\right\}\\
&\quad\times\left\{\left(\frac{1}{\sigma^2}\right)^{(N-K)/2+1}\exp\left(-\frac{(N-K)s^2}{2\sigma^2}\right)\right\}
\end{aligned}$$

给定 σ^2时 $\boldsymbol{\beta}$ 的条件后验分布 $p(\boldsymbol{\beta}|\sigma^2,\mathbf{y},\mathbf{X})$，而且数据 $\mathbf{y}$、$\mathbf{X}$ 显然都是 K 维多元变量正态的，其均值为 $\hat{\boldsymbol{\beta}}$ 且方差为 $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$，因为 $\boldsymbol{\beta}$ 仅仅出现在最终表达式的第一行。给定 $\boldsymbol{\beta}$ 时 σ^2的条件后验，由于 σ^2在表达式最终两行都出现了，所以它更难求出。

通过积分去掉 σ^2而求 $\boldsymbol{\beta}$ 的边缘后验，对于推断 $\boldsymbol{\beta}$ 的后验而言，这样做极为有用。我们对式(13.26)的第二行进行积分，做变量变换 $z=1/\sigma^2$，并运用给定常值

$a>0$、$c>-1$ 时 $\int_0^\infty z^c\exp(-az)dz=\Gamma(c+1)/a^{c+1}$ 的结果，这里 $c=N/2+1$，而 $a=\{\cdot\}$ 表示大括号中的长项。从而，得到边缘后验分布的核：

$$p(\boldsymbol{\beta}|\mathbf{y},\mathbf{X})\propto\{(N-K)s^2+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\}^{-N/2} \qquad (13.27)$$
$$\propto\{1+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'(s^2(N-K)(\mathbf{X}'\mathbf{X})^{-1})^{-1}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\}^{-(N-K+K)/2}$$

由 13.3.5 节知，这是在 $\hat{\boldsymbol{\beta}}$ 处中心化的多元变量学生 t 分布的核，其自由度为 $N-K$，而协方差矩阵 $s^2(\mathbf{X}'\mathbf{X})^{-1}$ 用 $(N-K)/(N-K-2)$ 乘。因而，有：

$$\boldsymbol{\beta}\sim t_K(\hat{\boldsymbol{\beta}},s^2(\mathbf{X}'\mathbf{X})^{-1}) \qquad (13.28)$$

$\boldsymbol{\beta}$ 的单个元素服从单变量学生 t 分布。

关于 σ^2 的边缘后验更容易获得，如果将式(13.26)中最终表达式对 β 进行积分，同时注意到，β 只在最终表达式第一行出现，这是 $\mathcal{N}[\hat{\boldsymbol{\beta}},\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]$ 密度的核且积分为 1。由此可得，σ^2 的边缘后验是：

$$p(\sigma^2|\mathbf{y},\mathbf{X})\propto(\sigma^2)^{-(N-K+1)/2}\exp\left(-\frac{(N-K)s^2}{2\sigma^2}\right) \qquad (13.29)$$

这个表达式是众所周知的反向平方根伽玛密度的核。也就是说，它是下述随机变量的密度，即含有自由度参数为 $N-K$ 的伽玛分布随机变量的平方根的倒数。该结果等同于频率学派在 $\hat{\boldsymbol{\beta}}$ 分布下得到的结果。

因此，对于正态线性回归来说，含有非信息先验的贝叶斯分析会产生在数量上类似于标准频率学派分析在有限样本中所获得的那些结论。以 σ^2 为条件的 $\boldsymbol{\beta}$ 后验服从 $\mathcal{N}[\hat{\boldsymbol{\beta}},\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]$ 分布，而无条件的 $\boldsymbol{\beta}$ 后验服从多元变量 t 分布。

可是，由于这些分布具有未知参数 $\boldsymbol{\beta}$ 的形式，且 $\boldsymbol{\beta}$ 具有均值 $\hat{\boldsymbol{\beta}}$ 而不是估计值 $\hat{\boldsymbol{\beta}}$ 的形式，对它们的解释截然不同。例如，关于 β_j 的贝叶斯 95% HPD 区间是 $\hat{\beta}_j\pm t_{0.025,N-K}\times\mathrm{se}[\hat{\beta}_j]$，其中，$\mathrm{se}[\hat{\beta}_j]=(s^2(\mathbf{X}'\mathbf{X})^{jj})^{1/2}$。由 13.2.5 节知，对此解释是 β_j 以后验概率 0.95 位于这个区间。

13.3.2 *信息先验*

如果我们使用关于 $\boldsymbol{\beta}$ 与 σ 的独立共轭，那么在信息先验下，对正态线性回归模型进行贝叶斯分析，特别有洞察力。由 13.2.4 节知，关于 $\boldsymbol{\beta}$ 的共轭先验是正态的，而关于 $1/\sigma^2$ 的共轭则是伽玛的。从而，得出正态—伽玛先验(**normal - gamma prior**)：

$$\pi(\boldsymbol{\beta},1/\sigma^2)=\pi_N(\boldsymbol{\beta}|1/\sigma^2)\pi_\gamma(1/\sigma^2)$$

其中，$\pi_N(\boldsymbol{\beta}|1/\sigma^2)$ 表示 $\mathcal{N}[\boldsymbol{\beta}_0,\sigma^2\boldsymbol{\Omega}_0^{-1}]$ 密度，$\boldsymbol{\beta}_0$ 与 $\boldsymbol{\Omega}_0$ 均为已知，而核为：

$$\pi_N(\boldsymbol{\beta}|1/\sigma^2)\propto\sigma^{-K}\exp\left[-\frac{(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{\Omega}_0(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}{2\sigma^2}\right] \qquad (13.30)$$

而 $\pi_\gamma(1/\sigma^2)$ 表示 $\mathcal{G}[\nu_0,s_0^2]$ 密度，其中，ν_0 与 s_0^2 均为已知常值，并且：

$$\pi_\gamma(1/\sigma^2)=\sigma^{-(\nu_0+1)}\exp\left[-\frac{\nu_0s_0^2}{2\sigma^2}\right] \qquad (13.31)$$

注意到，(局部)参数$\boldsymbol{\beta}$的先验依赖于(标度)参数σ。当σ反映了建立在y上的标度是度量的，从而应影响到$\boldsymbol{\beta}$时，这就会有意义。给定这个先验与式(13.25)中的似然函数，其后验密度具有正态伽玛类型。在经过一些代数运算后，它变成如下形式：

$$p(\boldsymbol{\beta},1/\sigma|\mathbf{y},\mathbf{X})\propto(\sigma^2)^{-N/2}\exp\left[-\frac{s^2(N-K)}{2\sigma^2}\right]\exp\left[-\frac{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})}{2\sigma^2}\right]$$
$$\times(\sigma^2)^{-K/2}\exp\left[-\frac{(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{\Omega}_0(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}{2\sigma^2}\right]$$
$$\times(\sigma^2)^{-(\nu_0/2)-1}\exp\left[-\frac{\nu_0 s_1^2}{2\sigma^2}\right]$$
$$\propto(\sigma^2)^{(\nu_0+N)/2-1}\exp\left[-\frac{s_1^2}{2\sigma^2}\right](\sigma^2)^{-K/2}$$
$$\times\exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})'\boldsymbol{\Omega}_1(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})\right] \tag{13.32}$$

其中，$\bar{\boldsymbol{\beta}}$与$\boldsymbol{\Omega}_1^{-1}$表示$\boldsymbol{\beta}$的后验均值与方差，而$s_1^2$表示$\sigma^2$的后验均值，它们被定义成：

$$\bar{\boldsymbol{\beta}}=(\boldsymbol{\Omega}_0+\mathbf{X}'\mathbf{X})^{-1}(\boldsymbol{\Omega}_0\boldsymbol{\beta}_0+\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}) \tag{13.33}$$
$$\boldsymbol{\Omega}_1=(\boldsymbol{\Omega}_0+\mathbf{X}'\mathbf{X})$$
$$s_1^2=s_0^2+\hat{\mathbf{u}}'\hat{\mathbf{u}}+(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})'[\boldsymbol{\Omega}_0^{-1}+(\mathbf{X}'\mathbf{X})^{-1}](\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})$$

后验均值$\bar{\boldsymbol{\beta}}$可通过利用“完全平方”矩阵形式来获得。特别地，给定$K\times1$维向量$\boldsymbol{\beta}$、$\bar{\boldsymbol{\beta}}$、$\boldsymbol{\beta}_0$和$\hat{\boldsymbol{\beta}}$以及$K\times K$阶对称方阵$\mathbf{A}$与$\mathbf{B}$，可以证明：

$$(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\mathbf{A}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\mathbf{B}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})$$
$$=(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})'(\mathbf{A}+\mathbf{B})(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})+(\boldsymbol{\beta}_0-\bar{\boldsymbol{\beta}})'\mathbf{AB}(\mathbf{A}+\mathbf{B})^{-1}(\boldsymbol{\beta}_0-\bar{\boldsymbol{\beta}})$$

其中，$\bar{\boldsymbol{\beta}}=(\mathbf{A}+\mathbf{B})^{-1}(\mathbf{A}\boldsymbol{\beta}_0+\mathbf{B}\hat{\boldsymbol{\beta}})$。

$\boldsymbol{\beta}$与σ^2的联合边缘后验具有相同的正态伽玛形式作为先验。

给定σ^2时，$\boldsymbol{\beta}$的条件后验具有均值$\bar{\boldsymbol{\beta}}$，即先验均值$\boldsymbol{\beta}_0$与样本均值$\hat{\boldsymbol{\beta}}$的加权矩阵平均。

通常，利用共轭先验在代数上等价于使用源自相同分布的样本来增加数据。在此情况下，正态—伽玛先验等价于满足下述条件的同样过程的额外样本，即该过程具有$\boldsymbol{\beta}_0$的回归参数估计值，$\mathbf{X}'\mathbf{X}$矩阵等于$\boldsymbol{\Omega}_0$，自由度参数等于ν_0，并且误差平方和等于$\nu_0 s_0^2$，由于$\boldsymbol{\Omega}_0$是一个固定矩阵，所以当$N\to\infty$时，$\boldsymbol{\Omega}_0/N\to\mathbf{0}$，而$\mathbf{X}'\mathbf{X}/N$收敛到常值矩阵。因此，若验证在大样本下，ML估计量与后验均值是等价的，则$\bar{\boldsymbol{\beta}}\to\hat{\boldsymbol{\beta}}$。后验方差$\boldsymbol{\Omega}_1^{-1}$与$(\boldsymbol{\Omega}_0+\mathbf{X}'\mathbf{X})^{-1}$成比例。更详细解释，参见利默(Leamer，1978)。

$\boldsymbol{\beta}$的边缘后验可通过对联合后验积分σ^2而获得。从而得到：

$$p(\boldsymbol{\beta}|\mathbf{y},\mathbf{X})\propto[s_1^2+(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})'(\boldsymbol{\Omega}_0+\mathbf{X}'\mathbf{X})(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})]^{-(\nu_1+K/2)} \tag{13.34}$$

因此，边缘后验是多元变量学生 t 分布，如同非信息先验情况一样，该分布是以 $\bar{\beta}$ 为中心的，而不是以 $\hat{\beta}$ 为中心的。

由于共轭先验处理先验信息时，就像前面源自同样过程的样本一样，所以即使来自两个来源的信息可能处于矛盾之中，但仍可对样本信息与先验信息进行对称研究。因而，利用共轭先验在数学形式的方便，无疑付出很高的代价。当先验信息与样本信息明显表现出矛盾时，可以预期后验分布具有双峰，其中一个峰值对应于样本均值，而另一个对应于先验均值。先验分布允许人们捕获这类特性，该先验分布意指设定 β 具有独立于 $1/\sigma^2$ 的多元变量学生 t 密度，而 $1/\sigma^2$ 具有独立于 $\mathbf{X}\beta$ 的伽玛先验分布。这被称为"迪基先验"(**Dickey's prior**)[利默(Leamer, 1978,第 79 页)]。在此假设下，边缘后验是两个多元变量学生 t 密度之积；该积也可表述成两个 t 分布的混合。这种分布能潜在揭示出两峰性。利默(Leamer, 1978)对这种情况曾经给出更为广泛的分析。

13.3.3 *混合估计*

在频率学派背景下，我们探索线性回归贝叶斯分析的用武之地。

通常，频率学派分析将先验信息并入等式约束之中，这是先验中的方差参数趋于零的贝叶斯分析的限制情形。相反，作为随机的先验信息也可被并入频率学派分析中，只是要利用混合估计(**mixed estimation**)。这种代数运算简单，并且该方法还提供一种方法以直观认识贝叶斯方法是如何将先验信息和样本信息融合在一起的。

在正态性下，我们继续研究线性回归模型。假定回归参数的先验信息 $\beta\sim\mathcal{N}[\mathbf{0},\ \sigma_v^2\mathbf{I}_K]$，这一点相对很容易地推广到非零均值上。将先验信息写成：

$$\beta=\mathbf{0}+\mathbf{v}$$

其中，$\mathbf{v}$ 表示 $K\times1$ 维误差，满足 $\mathbf{v}\sim\mathcal{N}[\mathbf{0},\ \sigma_v^2\mathbf{I}_K]$。现在，通过这个先验对 $\mathbf{y}=\mathbf{X}\beta+\mathbf{u}$ 样本信息扩大，并把整个模型写成增广回归模型(**augmented regression model**)：

$$\begin{bmatrix}\mathbf{y}\\ \mathbf{0}\end{bmatrix}=\begin{bmatrix}\mathbf{X}\\ \mathbf{I}_K\end{bmatrix}\beta+\begin{bmatrix}\mathbf{u}\\ -\mathbf{v}\end{bmatrix}$$

经过重新参数化，得到：

$$\begin{aligned}\begin{bmatrix}\mathbf{y}\\ \mathbf{0}\end{bmatrix}&=\begin{bmatrix}\mathbf{X}\\ \dfrac{\sigma}{\sigma_v}\mathbf{I}_K\end{bmatrix}\beta+\begin{bmatrix}\mathbf{u}\\ -\dfrac{\sigma}{\sigma_v}\mathbf{v}\end{bmatrix}\\ &=\begin{bmatrix}\mathbf{X}\\ \lambda\mathbf{I}_K\end{bmatrix}\beta+\begin{bmatrix}\mathbf{u}\\ \mathbf{v}^*\end{bmatrix}\end{aligned}\tag{13.35}$$

其中，$\lambda=\sigma/\sigma_v$，使用了变换 $\mathbf{v}^*=-\lambda\mathbf{v}$，因此，所有误差具有共同方差 σ^2。

建立在这种增大数据集基础上的估计量是合并估计量(**pooled estimator**)或混合估计量(**mixed estimator**)。以 λ 为条件的混合估计量是：

$$\begin{aligned}\hat{\beta}_\lambda &= [\mathbf{X}'\mathbf{X}+\lambda^2\mathbf{I}_k]^{-1}\mathbf{X}'\mathbf{y} \\ &= [\mathbf{X}'\mathbf{X}(\mathbf{I}_k+\lambda^2(\mathbf{X}'\mathbf{X})^{-1})]^{-1}\mathbf{X}'\mathbf{y} \\ &= [\mathbf{I}_k+\lambda^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \mathbf{A}_\lambda\hat{\beta}\end{aligned} \quad (13.36)$$

其中，$\mathbf{A}_\lambda=[\mathbf{I}_k+\lambda^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}$，而$\hat{\beta}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$表示无约束 OLS 估计量。

这个估计量是所谓的岭回归估计量，由霍尔和肯纳德(Hoerl and Kennard, 1970)在无贝叶斯分析理由的情况下引入，以此对抗小样本的多重共线性问题。该估计量还归属于压缩估计量[1](**shrinkage estimator**)，此估计量压缩到(或被拉向)先验均值，在此情况下，即压缩到零向量。有时，在有限样本拥有多重共线性数据时，这样做就有意义，其中"t 比率"趋于 0，在变量系数真实趋于 0 与那些变量系数仅仅看来好像是 0 之间进行辨别很难。在极限形式下，压缩将变量排除在外。

值得注意$\hat{\beta}_\lambda$的几个特性：(1)以λ为条件的$\hat{\beta}_\lambda$表示β后验分布的均值；(2)此估计量是 **0** 向量与$\hat{\beta}_\lambda$的矩阵加权平均(**matrix-weighted average**)；(3)如果我们选取使估计量向着某个非零β收敛，比如说β_0，那么代数运算几乎没有什么变化。于是，所得到的估计量是向量β_0与$\hat{\beta}$的矩阵加权平均(**matrix-weighted average of vectors**)。

当$N\to\infty$时，对称加权矩阵$\mathbf{A}_\lambda=[\mathbf{I}_k+(\lambda^2/N)(N^{-1}\mathbf{X}'\mathbf{X})^{-1}]\to\mathbf{I}_K$，这是因为$\lambda^2/N\to\mathbf{0}$。因此：

$$\text{当 } N\to\infty\text{ 时，}\ \hat{\beta}_\lambda\to\hat{\beta}$$

所以先验对后验均值的影响会随着样本变大而消失。类似地，$\hat{\beta}_\lambda$的条件后验方差由：

$$\begin{aligned}\mathrm{V}[\hat{\beta}_\lambda] &= \mathbf{A}_\lambda\mathrm{V}[\hat{\beta}]\mathbf{A}_\lambda \\ &= \sigma^2\mathbf{A}_\lambda(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}_\lambda\end{aligned}$$

给出，所以当样本量$N\to\infty$时，$\mathrm{V}[\hat{\beta}_\lambda]\to\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$。

对于有限样本来说，以λ与σ^2为条件，$\hat{\beta}_\lambda$的条件后验分布(**posterior distribution**)为：

$$\hat{\beta}_\lambda|\lambda,\ \sigma^2\sim\mathcal{N}[\mathbf{A}_\lambda\hat{\beta},\ \sigma^2\mathbf{A}_\lambda(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}_\lambda'] \quad (13.37)$$

$\hat{\beta}_\lambda$的边缘后验分布可通过积分去掉λ与σ^2而获得。若将λ处理成给定的，且假定关于σ^2具有不明确先验或非信息先验，就能积分去掉σ^2，正如 13.3.1 节所证明的。这种积分运算在解析形式上是可行的，而且会得到β_λ的边缘后验，它是多元变量学生 t 分布的。最后，我们设定λ的先验分布，由于$\lambda>0$，即可能的伽玛先验，然后继续积分去掉它。不过，λ以难以处理的方式进入条件后验中，而且在解析形式上不能积分去掉它。在此情况下，需要采用数值技术。假定可以这样实施，然后对这个模型进行贝叶斯处理。

---

[1] 又称为压缩型估计量。——译者注

13.3.4 分层先验

我们考察三阶段线性回归模型，该模型关于回归参数是分层的，而方差参数则不是。

第一阶段是线性回归模型，记为 $\mathbf{y}=\mathbf{X}_1\boldsymbol{\beta}_1+\mathbf{u}$，其中，增加下标 1 用以区分第一阶段的参数及回归元与第二阶段的参数及回归元。参数 $\boldsymbol{\beta}_1$ 是随机的，并且对其建模既依赖于参数，又依赖于数据，因此 $\boldsymbol{\beta}_1=\mathbf{X}_2\boldsymbol{\beta}_2+\mathbf{v}$。例如，第一层对个体学生检验成绩建模，而第二层则对学校特性建模。假定误差是正态分布的。第二阶参数 $\boldsymbol{\beta}_2$ 被处理成未知的，并对它设定一个先验。同理，对第一阶段模型中的方差参数 σ_1^2 加以设定。

假定正态分布误差，并且利用共轭先验，会产生下述模型：

$$\mathbf{y}|\mathbf{X}_1,\ \boldsymbol{\beta}_1,\ \sigma_1^2\sim\mathcal{N}[\mathbf{X}_1\boldsymbol{\beta}_1,\ \sigma_1^2\mathbf{I}_N] \tag{13.38}$$

$$\boldsymbol{\beta}_1|\mathbf{X}_2,\ \boldsymbol{\beta}_2,\ \boldsymbol{\Sigma}_2\sim\mathcal{N}[\mathbf{X}_2\boldsymbol{\beta}_2,\ \boldsymbol{\Sigma}_2] \tag{13.39}$$

$$\boldsymbol{\beta}_2\sim\mathcal{N}[\boldsymbol{\beta}^*,\ \boldsymbol{\Sigma}^*] \tag{13.40}$$

$$\sigma_1^{-2}|\nu^*,\ \sigma^{*2}\sim\mathcal{N}[\nu^*/2,\ \nu^*\sigma^{*2}/2] \tag{13.41}$$

其中，$\mathbf{X}_1$ 表示 $N\times K$ 的，$\mathbf{X}_2$ 表示 $K\times M$ 的，$\boldsymbol{\beta}_1$ 表示 $K\times1$ 的，$\boldsymbol{\beta}_2$ 表示 $M\times1$ 的，$\boldsymbol{\Sigma}_2$ 表示 $K\times K$ 的，$\boldsymbol{\beta}^*$ 表示 $M\times1$ 的，而 $\boldsymbol{\Sigma}^*$ 表示 $M\times M$ 的。对于回归参数 $\boldsymbol{\beta}_1$ 来说，第二行给出其先验，而第三行给出后面第二阶段关于 $\boldsymbol{\beta}_2$ 的先验，或者先验之先验（尽管 $\boldsymbol{\Sigma}_2$ 被假定是已知的），参数（$\boldsymbol{\beta}^*$，$\boldsymbol{\Sigma}^*$）常常称为超参数。对于方差参数来说，第四行给出方差参数 σ_1^2 的先验，ν^* 与 σ^{*2} 是已经设定的。新的内容是增加部分[式(13.40)]。

注意到，我们将一些阶段叠放起来。而且，将这转变成两层模型。特别地，利用信息先验两种方式之一，可写出两阶段模型，或者：

$$\mathbf{y}|\mathbf{X}_1,\ \boldsymbol{\beta}_1,\ \sigma_1^2\sim\mathcal{N}[\mathbf{X}_1\boldsymbol{\beta}_1,\ \sigma_1^2\mathbf{I}_N]$$
$$\boldsymbol{\beta}_1|\mathbf{X}_2,\ \boldsymbol{\Sigma}_2\sim\mathcal{N}[\mathbf{X}_2\boldsymbol{\beta}^*,\ \boldsymbol{\Sigma}_2+\mathbf{X}_2\boldsymbol{\Sigma}^*\mathbf{X}_2']$$

或者：

$$\mathbf{y}|\mathbf{X}_1,\ \mathbf{X}_2,\ \boldsymbol{\beta}_2,\ \boldsymbol{\Sigma}_2,\ \sigma_1^2\sim\mathcal{N}[\mathbf{X}_1\mathbf{X}_2\boldsymbol{\beta}_2,\ \sigma_1^2\mathbf{I}_N+\mathbf{X}_1\boldsymbol{\Sigma}_2\mathbf{X}_1']$$
$$\boldsymbol{\beta}_2\sim\mathcal{N}[\boldsymbol{\beta}^*,\ \boldsymbol{\Sigma}^*]$$

若 σ_1^2 给定，这种设置背景对应于条件共轭（**conditionally conjugate**）正态先验。利用前面介绍的结果，将 $\boldsymbol{\beta}_1$ 或者 $\boldsymbol{\beta}_2$ 的后验均值表达式推导成为 $\boldsymbol{\beta}^*$ 与 $\hat{\boldsymbol{\beta}}_1$ 或 $\boldsymbol{\beta}^*$ 与 $\hat{\boldsymbol{\beta}}_2$ 的矩阵加权平均。

运用正态分布只是一种阐述性的。关于广义线性模型的分层模型，即线性指数族的成员，具有广泛应用[阿尔伯特(Albert, 1988)]。

在分层模型中，以便于处理的解析形式获得第一阶段参数比如 $\boldsymbol{\beta}_1$ 的后验概率分布是不可能的。幸运的是，下一节将要阐述的计算方法，特别适合于对层结构进行建模。

另一种方法，即对经验贝叶斯(**empirical Bayes**)的应用，涉及较高阶段先验中的参数估计，这类似于似然方法。例如，该方法避开假定 $\boldsymbol{\Sigma}_2$ 与 $\boldsymbol{\Sigma}^*$ 均是已知矩阵。

13.3.5 多元变量 t 分布与威沙特分布

与经典分析相比，贝叶斯分析使用更广泛的分布。这里，对线性回归在正态性下贝叶斯分析用到的两个多元变量分布加以详细阐述。

多元变量 t 分布是将单变量学生 t 分布推广到多元变量的情形。它类似于多元变量正态分布，只是其分布尾部相当宽。在贝叶斯分析中，给出关于 $\boldsymbol{\beta}$ 的后验分布，共轭正态先验(参见 13.3.2 节)或能直接用作关于 $\boldsymbol{\beta}$ 的先验，当其尾部比人们期望的正态尾部大时，便出现多元变量 t 分布。一个 $q\times 1$ 维随机变量 t 作为多元变量学生 t 分布，其中，自由度参数为 υ，均值为 μ 且分散参数为 $\boldsymbol{\Sigma}$，它具有联合密度：

$$f_t(\mathbf{t}|\upsilon,\ \boldsymbol{\mu},\ \Sigma)=\frac{\Gamma((\upsilon+1)/2)}{\Gamma(\upsilon/2)(\pi\upsilon)^{(1/2)}|\Sigma|^{1/2}}\times\left\{1+\frac{1}{\upsilon}(\mathbf{t}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{t}-\boldsymbol{\mu})\right\}^{-(\upsilon+q)/2}$$

其中，$\Gamma(\cdot)$表示伽玛函数。这个分布关于众数 $\boldsymbol{\mu}$ 是对称的，当 $\upsilon>1$ 时，均值为 $\boldsymbol{\mu}$，而当 $\upsilon>2$ 时，方差为$[\upsilon/(\upsilon-2)]\Sigma$。其尾部比正态的要宽一些(例如，若 $\upsilon=3$，则方差为 3Σ)，提出一种容易获得抽样的方式，同时当 $\upsilon\to\infty$时，变成正态情况。若 $\mathbf{z}\sim\mathcal{N}[\mathbf{0},\ \mathbf{I}]$，且 $s\sim\chi^2(\upsilon)$ ，则 $\mathbf{t}=\boldsymbol{\mu}+\Sigma^{-1/2}\mathbf{z}/\sqrt{s/\upsilon}$ 服从此处给出的多元变量 t 分布。

威沙特分布是单变量卡方分布推广到多元变量的情形，或更一般的伽玛分布。在贝叶斯分析中，它用作多元变量正态分布的协方差矩阵逆的共轭先验。一个 $q\times q$ 阶随机正定矩阵 $\mathbf{W}$ 作为威沙特分布(**Wishart distributed**)，其自由度参数 $\upsilon\geqslant q$，且标度矩阵 $\mathbf{S}$，它具有联合密度：

$$f_{\mathrm{W}}(\mathbf{W}|\upsilon,\ \mathbf{S})=2^{\upsilon q/2}\pi^{q(q-1)/4}\prod_{j=1}^{q}\Gamma\left(\frac{\upsilon+1-j}{2}\right)\times|\mathbf{S}|^{-\upsilon/2}|\mathbf{W}|^{(\upsilon-q-1)/2}\exp(-\operatorname{tr}(\mathbf{S}^{-1}\mathbf{W})/2)$$

其中，$\Gamma(\cdot)$表示伽玛函数，而 $\operatorname{tr}(\cdot)$表示矩阵的迹。这一分布具有均值 $\upsilon\mathbf{S}$。关于 iid 多元变量正态数据的样本协方差矩阵就是威沙特分布。更一般地，给定 $\upsilon(q)$，独立的 $q\times 1$ 维向量 $\mathbf{x}_j\sim\mathcal{N}[\mathbf{0},\ \mathbf{S}]$，则 $\sum_{j=1}^{\upsilon}\mathbf{x}_j\mathbf{x}_j'$ 服从威沙特分布。当 $\mathbf{W}^{-1}$服从密度为 $f_{\mathrm{W}}(\mathbf{W}^{-1}|\upsilon,\ \mathbf{S})$的威沙特分布时，$\mathbf{W}$ 服从逆威沙特分布(**inverse-Wishart distributed**)，其密度为：

$$f_{\mathrm{IW}}(\mathbf{W}|\upsilon,\ \mathbf{S})=2^{\upsilon q/2}\pi^{q(q-1)/4}\prod_{j=1}^{q}\Gamma\left(\frac{\upsilon+1-j}{2}\right)|\mathbf{S}|^{\upsilon/2}|\mathbf{W}|^{-(\upsilon+q+1)/2}\exp(-\operatorname{tr}(\mathbf{S}^{-1}\mathbf{W})/2)$$

13.4 蒙特卡罗积分

在许多建模情况中，关注参数的后验分布在解析形式上是难以处理的。在这类情况下，需要数值方法来估计全部后验分布，或者估计该分布的某个重要成分，诸如后验均值。

在本节，我们考察在没有以显性方式获得后验分布时对重要后验矩进行计算。第 12 章的一些方法能得以应用，只需潜在而很少的计算，对于整个样本而不是每个个体都要进行每次迭代。在下一节，将阐述模拟后验分布的方法。

13.4.1 *重要抽样*

假如问题是计算后验矩函数 $\mathrm{E}[m(\boldsymbol{\theta}|\mathbf{y})]$，其中，期望是关于后验密度 $p(\boldsymbol{\theta}|\mathbf{y})$ 的。我们想要计算：

$$\mathrm{E}[m(\boldsymbol{\theta})]=\int_{R(\boldsymbol{\theta})} m(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \tag{13.42}$$

例如，第 k 个参数的后验均值是 $\mathrm{E}[\theta_k]=\int\theta_k p(\boldsymbol{\theta}\mid\mathbf{y})d\boldsymbol{\theta}$。其他一些例子包括后验标准差、边缘后验密度、后验区间以及给定参数函数的后验期望。

由第 12 章知，对 $\mathrm{E}[m(\boldsymbol{\theta})]$ 的直接蒙特卡罗估计是 $\hat{\mathrm{E}}[m(\boldsymbol{\theta})]=S^{-1}\sum_s m(\boldsymbol{\theta}^s)$，其中，$\boldsymbol{\theta}^s$，$s=1,\cdots,S$ 都是从后验密度 $p(\boldsymbol{\theta}|\mathbf{y})$ 中得到的 S 个 $\boldsymbol{\theta}$ 采样。不过，在目前贝叶斯背景下，当式(13.1)中正式定义的后验密度不存在闭形式解，这种估计行不通，进而不可能从后验 $p(\boldsymbol{\theta}|\mathbf{y})$ 中实施采样。可是，我们能使用 12.7.2 节已经引进的重要抽样。所考察的式(13.42)积分重新写成：

$$\mathrm{E}[m(\boldsymbol{\theta})]=\int_{R(\boldsymbol{\theta})}\left(\frac{m(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})}{g(\boldsymbol{\theta})}\right)g(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{13.43}$$

其中，$g(\boldsymbol{\theta})>0$ 表示已知的密度函数，它与 $p(\boldsymbol{\theta}|\mathbf{y})$ 具有相同支集，这很容易进行采样。其对应的蒙特卡罗积分估计是：

$$\hat{\mathrm{E}}[m(\boldsymbol{\theta})]=\frac{1}{S}\sum_{s=1}^{S}\frac{m(\boldsymbol{\theta}^s)p(\boldsymbol{\theta}^s|\mathbf{y})}{g(\boldsymbol{\theta}^s)}$$

其中，$\boldsymbol{\theta}^s$ 表示从重要抽样密度(**importance sampling density**) $g(\boldsymbol{\theta})$ 而不是从最初目标密度(**target density**) $p(\boldsymbol{\theta}|\mathbf{y})$ 中得到的 S 个 $\boldsymbol{\theta}$ 采样，$s=1,\cdots,S$。注意到，如果 $p(\boldsymbol{\theta}|\mathbf{y})$ 依赖于额外的参数，或者完全条件密度的函数形式是已知的，但边缘后验的函数形式是未知的，要求 $p(\boldsymbol{\theta}|\mathbf{y})$ 与 $g(\boldsymbol{\theta})$ 应该具有相同的支集会出现潜在问题。

此外，应用后验密度需要解释式(13.1)分母中的积分常值。设 $p^{\mathrm{ker}}(\boldsymbol{\theta}|\mathbf{y})$ 表示后验密度的核(**kernel**)，其中，$p^{\mathrm{ker}}(\boldsymbol{\theta}|\mathbf{y})=\mathrm{L}(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$，或者是这个量的倍数。然而，为了记号简单起见，在下文中不使用关于 $\mathbf{y}$ 的依赖性。于是，后验密度为：

$$p(\boldsymbol{\theta})=\frac{p^{\mathrm{ker}}(\boldsymbol{\theta})}{\int p^{\mathrm{ker}}(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

其对应的后验矩为：

$$
\begin{aligned}
\mathrm{E}[m(\boldsymbol{\theta})] &= \int m(\boldsymbol{\theta})\left[\frac{p^{\mathrm{ker}}(\boldsymbol{\theta})}{\int p^{\mathrm{ker}}(\boldsymbol{\theta})d\boldsymbol{\theta}}\right]d\boldsymbol{\theta} \\
&= \frac{\int m(\boldsymbol{\theta})p^{\mathrm{ker}}(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int p^{\mathrm{ker}}(\boldsymbol{\theta})d\boldsymbol{\theta}} \\
&= \frac{\int (m(\boldsymbol{\theta})p^{\mathrm{ker}}(\boldsymbol{\theta})/g(\boldsymbol{\theta}))g(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int (p^{\mathrm{ker}}(\boldsymbol{\theta})/g(\boldsymbol{\theta}))g(\boldsymbol{\theta})d\boldsymbol{\theta}}
\end{aligned}
$$

于是，后验矩 $\mathrm{E}[m(\boldsymbol{\theta})]$ 的重要基于抽样估计（**important sampling-based estimate**）是：

$$
\hat{\mathrm{E}}[m(\boldsymbol{\theta})] = \frac{\frac{1}{S}\sum_{s=1}^{S} m(\boldsymbol{\theta}^s)p^{\mathrm{ker}}(\boldsymbol{\theta}^s)/g(\boldsymbol{\theta}^s)}{\frac{1}{S}\sum_{s=1}^{S} p^{\mathrm{ker}}(\boldsymbol{\theta}^s)/g(\boldsymbol{\theta}^s)} \tag{13.44}
$$

其中，$\boldsymbol{\theta}^s$ 表示从重要抽样密度 $g(\boldsymbol{\theta})$ 中得到的 S 个 $\boldsymbol{\theta}$ 采样，$s=1,\cdots,S$。

这个方法是由克洛克和范迪克（Kloek and Van Dijk, 1978）提出的。在某些正则条件下，格韦克（Geweke, 1989）建立了一致性与渐近正态性。这些条件包括下述几个假设：在 $p(\boldsymbol{\theta})$ 的支集 $R(\boldsymbol{\theta})$ 上，重要抽样密度 $g(\boldsymbol{\theta})>0$；$\mathrm{E}[m(\boldsymbol{\theta})]<\infty$，因而后验矩存在；$\int p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}=1$，从而后验密度是正常的。正如前面已注意的，通常我们以核 $p^{\mathrm{ker}}(\boldsymbol{\theta}|\mathbf{y})=\mathrm{L}(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ 来进行分析，这不必积分为 1。先验 $\pi(\boldsymbol{\theta})$ 不必是正常的，但为了确保 $\int p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}=1$，它必须满足 $\int \pi(\boldsymbol{\theta})d\boldsymbol{\theta}<\infty$。

重要抽样方法虽然简单，但格韦克（Geweke, 1989）指出，实施起来需要给出良好的巧妙解释。一个关键性要求是，$g(\boldsymbol{\theta})$ 应比 $p(\boldsymbol{\theta}|\mathbf{y})$ 具有更厚的尾部，以此确保重要权数（**importance weight**）$w(\boldsymbol{\theta})=p(\boldsymbol{\theta}|\mathbf{y})/g(\boldsymbol{\theta})$ 有界。鉴于后验对数的渐近正态性，对 $g(\boldsymbol{\theta})$ 的良好选择是多元变量 t 分布，对其均值设定为后验众数，而协方差矩阵与后验对数海赛阵的逆成正比，同时对自由度设定为充分小的值，以便保证其原尾部。格韦克（Geweke, 1989）还提供了所谓的数值有效性（**numerical efficiency**），估计利用从 $g(\boldsymbol{\theta})$ 得到的抽样进行计算来达到 $\hat{\mathrm{E}}[m(\boldsymbol{\theta})]$ 给定准确性水平所需要的复制次数，相对于从 $p(\boldsymbol{\theta}|\mathbf{y})$ 进行采样可能所需要的复制次数。由第 12 章知，对于较高阶积分，为了获得积分良好近似就需要更多次的模拟采样，而且人们可另外使用第 12 章曾阐述的模拟加速法，例如对偶抽样。

重要抽样方法使用了等概率从抽样密度 $g(\boldsymbol{\theta})$ 中得到的采样。一种更有效的近似可以依照 $g(\boldsymbol{\theta}^s)$ 接近目标 $p(\boldsymbol{\theta}^s|\mathbf{y})$ 的程度，来对采样给予权数。这通过重要再抽样来完成[参见格尔曼（Gelman, 1995）]。

重要抽样方法可用于提供后验的许多有用概括性测量，如同 13.2.5 节所阐述的。允许计算 95%后验区间与 θ_k 后验密度图，这就会包括后验的分位数及百分位数估计值。

13.5 马尔可夫链蒙特卡罗模拟

贝叶斯分析的现代思想更加专注于对后验分布的重要概括性测量进行估计(参见前面一节),因为从后验分布获得大样本是人们所期望的。然后,来自后验的这种样本概括统计量将会提供所期望的有关估计值的样本矩特征信息以及其他有意思的相关测量信息,比如参数的边缘分布或者参数函数信息。例如,给定从后验分布中得到的 S 个采样,通过 $S^{-1}\sum_s\theta_k^s$ 估计 $\mathrm{E}[\theta_k]$。

当后验密度不存在容易处理的闭形式表达式时,挑战是从联合后验分布中获得采样。如果利用重要抽样对后验矩计算,存在适当密度,那么利用 12.8 节所述的筛选法从后验中采样同样是合适的。不过,当出现拒绝高百分位数时,该方法便无效。

然而,序贯采样(**sequential draws**)会使得产生的模拟值收敛到平稳分布上。如果实施的序贯是足够长的,该平稳分布与目标后验密度 $p(\boldsymbol{\theta}|\mathbf{y})$ 相一致。这种方法称为马尔可夫链蒙特卡罗(**Markov chain Monte Carlo, MCMC**),因为它涉及(蒙特卡罗)模拟且序列是马尔可夫形式的。在此链收敛之后,S 个序贯采样用于计算后验的概括性测量,比如通过 $\hat{\mathrm{E}}[\theta_k]=S^{-1}\sum_s\theta_k^s$ 估计 $\mathrm{E}[\theta_k]$。不过,一些采样是正相关的,故对于给定的 S 来说,估计的准确性将会减少,因为其估计方差将大于通常的 $(S-1)^{-1}\sum_s(\theta_k^s-\hat{\mathrm{E}}[\theta_k])^2$。

序贯方法要求构造马尔可夫链。两种广泛使用的计算法是,吉布斯抽样器(**Gibbs sampler**)与梅特罗波利斯—黑斯廷斯(**Metropolis - Hastings**)算法,前者是后者的一种特殊情况,参见黑斯廷斯(Hastings, 1970)。针对该主题的详细研究内容,可在格尔曼等人(Gelman et al., 1995)、盖默曼(Gamerman, 1997)以及罗伯特和卡塞拉(Robert and Casella, 1999)的文献中找到。下述内容是一个基本概述。

13.5.1 *马尔可夫链*

在阐述吉布斯抽样器与梅特罗波利斯—黑斯廷斯算法之前,我们给出 MCMC 文献中使用的一些重要定义及概念。这些定义是在离散状态模型背景下给出的。可将它们推广到连续状态模型,用于后验关于参数为连续的有关应用。

马尔可夫链(**Markov chain**)被定义成随机变量 $x_n(n=0,1,2,\cdots)$ 的一个序列,其中,x_n 在有限空间 A 中取值,并且把 x_n 定义成等于给定前面一些 x_{n-j} 值时特殊值转移核(**transition kernel**)。考察具有下述性质:

$$\Pr[x_{n+1}=x|x_n,x_{n-1},\cdots,x_0]=\Pr[x_{n+1}=x|x_n] \tag{13.45}$$

的马尔可夫链,因而给定过去 x_{n+1} 的分布仅仅由前面值 x_n 完全决定。这种转移核是一个转移矩阵 $\mathbf{T}$ (**transition matrix**),它的元素满足:

$$t_{xy}=\Pr[x_{n+1}=y|x_n=x] \tag{13.46}$$

非正式地讲,它表示从 x 到 y 的概率。对于有限状态(**finite-state**)马尔可夫链来说,x_n 可能取值的集合 A (状态)是具有有限个元素,比如说 m 个。于是:

$$\mathbf{T}=\begin{bmatrix} t_{11} & \cdots & t_{1m} \\ \vdots & \ddots & \vdots \\ t_{m1} & \cdots & t_{mm} \end{bmatrix} \tag{13.47}$$

其中，$\sum_{j=1}^{m} t_{ij}=1,\ i=1,\cdots,m$。

现在，考察从 x 到 y 用了 n 步（阶段）的转移。该转移概率由 $\mathbf{T}^n$ 给出，即 $\mathbf{T}$ 的 n 次矩阵积。矩阵 $\mathbf{T}^n$ 的行给出在第 n 阶段跨越 m 个状态的边缘分布，而第 j 列向量 $\mathbf{t}_j^{(n)}=(t_{j1}^{(n)},\cdots,t_{jm}^{(n)})$ 给出在第 n 阶段从状态 j 到其他状态的转移概率的边缘分布。若转移概率的初始分布记为 $\mathbf{t}_j^{(0)}$，则 $\mathbf{t}_j^{(n)}=\mathbf{t}_j^{(0)}\mathbf{T}^n=\mathbf{t}_j^{(n-1)}\mathbf{T}$。因此，在第 n 阶段上转移概率的边缘分布只是由初始分布与转移矩阵确定。在马尔可夫模拟背景下，当 $n\rightarrow\infty$ 时，链的渐近特性成为关注的内容。一个链称为可产生含有转移概率 t_{xy} 的平稳分布（**stationary distribution**）或不变分布（**invariant distribution**），如果：

$$\sum_{x\in A}\mathbf{t}_x\mathbf{T}_{x,y}=\mathbf{t}_y,\quad \forall y\in A \tag{13.48}$$

其中，转移是从状态 $\mathbf{t}_x$ 到 $\mathbf{t}_y$。然后，利用转移矩阵，从而得出转移概率的边缘分布没有任何变化。平稳分布的存在性与唯一性是一个重要问题。

若平稳分布存在，且 $\lim_{n\rightarrow\infty}\mathbf{t}_x\mathbf{T}_{x,y}^n=\mathbf{t}_y$，则此链与初始分布独立、渐近地趋向于 $\mathbf{t}_y$。在这个意义上，$\mathbf{t}_y$ 成为极限。尽管此处平稳分布是对有限状态马尔可夫链来定义的，但 MCMC 方法能处理马尔可夫链是无限状态的情况，参见吉尔克斯、理查森和施皮格尔霍尔特(Gilks，Richardson，and Spiegelhalter，1996，第 60～61 页)。

一个状态 y 可能是循环的或非常返状态的。一个循环状态（**recurrent state**）是指该状态以概率 1 重新返回，而一个非常返状态（**transient state**）是指该状态不会以某个正概率重新返回。

对于贝叶斯应用来说，目标是从后验 $p(\boldsymbol{\theta})$ 中获得采样。一旦应用马尔可夫链获得这些采样，参数向量的初始值 $\boldsymbol{\theta}^{(0)}$（它类似于状态的分布）是被指派的或者从转移核中抽样来的。若利用合适采样伪随机数的方法，则新向量值 $\boldsymbol{\theta}^{(1)}$ 可从在 $\boldsymbol{\theta}^{(0)}$ 处计算的转移核中采样，即 $K(\boldsymbol{\theta}^{(0)})$。在第 n 阶段中，采样是从转移核 $K(\boldsymbol{\theta}^{(n-1)})$ 中抽样等。所用的马尔可夫链使得当 $n\rightarrow\infty$ 时，极限分布成为后验 $p(\boldsymbol{\theta})$。一旦出现收敛到极限分布，所有序列采样也可以来自此分布，尽管采样序列将是相关的。

这些思想提供了 MCMC 程序类型的直观基础，而 MCMC 程序能用于从各种各样的可能高维数的模型中重新获得贝叶斯后验分布，例如在 13.3.4 节曾经讨论的线性分层模型。倘若人们设定从 θ 中采样而来的转移核 $K(\boldsymbol{\theta}^{(n-1)},\cdot)$，以及在其内可嵌入链的极限分布，则目标后验分布在任意紧密接近的情况下能够重新获得。

目前表述是在相当一般水平上给出的。实践中，对转移核的选择不是唯一的，并存在许多可能的人们能构造出的链。依照收敛到极限分布的速度来看，某些选择或许比其他一些要好。当人们发现收敛非常慢且计算量巨大时，就需要用可供选择的链来代替。很明显，链处于第 n 阶段时，需要一些准则来确定收敛是否出现以及接近到目标分布的程序。

13.5.2 吉布斯抽样器

我们以吉布斯抽样器[1](sampler)开始讨论，吉布斯抽样器作为 MCMC 类型的成员之一，容易对它给出描述并实施。

设 $\boldsymbol{\theta}=[\boldsymbol{\theta}_1\ \boldsymbol{\theta}_2]'$ 具有后验密度 $p(\boldsymbol{\theta})=p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$，这里为了记号简单起见，无须对 y 相依的记号。如果条件密度已知，就无须 $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)$ 和 $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)$ 的知识，那么可供选择的序列采样来自依极限收敛到从 $p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ 中得到的采样 $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)$ 与 $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)$。

例子

一种简单的阐明是考察具有均匀先验的均值及已知协方差矩阵的二元正态数据。设 $\mathbf{y}=(y_1, y_2)\sim\mathcal{N}[\boldsymbol{\theta},\boldsymbol{\Sigma}]$，其中，$\boldsymbol{\theta}=[\boldsymbol{\theta}_1\ \boldsymbol{\theta}_2]'$，而 $\boldsymbol{\Sigma}$ 具有对角元为 1 且非对角线元为 ρ。然后，给定关于 $\boldsymbol{\theta}$ 的均匀先验，可以证明，其后验是 $\boldsymbol{\theta}|\mathbf{y}\sim\mathcal{N}[\mathbf{y}, N^{-1}\boldsymbol{\Sigma}]$ 二元正态分布。由于条件后验分布是：

$$\theta_1|\theta_2, \mathbf{y}\sim\mathcal{N}[\bar{y}_1+\rho(\theta_2-\bar{y}_2), (1-\rho^2)/N]$$
$$\theta_2|\theta_1, \mathbf{y}\sim\mathcal{N}[\bar{y}_2+\rho(\theta_1-\bar{y}_1), (1-\rho^2)/N]$$

我们能够利用 θ_1 与 θ_2 的更新值从每一个条件正态分布中进行迭代抽样。若链实施足够长，则它将收敛到二元正态分布。在这个例子中，利用 12.8 节给出的乔列斯基变换，很容易地从 $\boldsymbol{\theta}|\mathbf{y}$ 联合后验中做出直接采样，可是在其他一些例子中，它可能从条件后验而不是联合后验中采样。

吉布斯抽样器

更一般地，考察 q 维目标分布 $p(\boldsymbol{\theta})$，这里不使用对数据相依的记号。假设 $\boldsymbol{\theta}$ 被分割成 d 个块。例如，线性回归例子的 $\boldsymbol{\theta}'=[\beta\sigma^2]'$。设 $\boldsymbol{\theta}_k$ 表示第 k 个块，而 $\boldsymbol{\theta}_{-k}$ 表示剔除 $\boldsymbol{\theta}_k$ 之后剩下的 $\boldsymbol{\theta}$ 成分。假定完全条件分布 $p(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{-k})$ 都是已知的，$k=1,\cdots,d$。于是，从完全条件中进行序贯抽样建立如下：

(1) 设 $\boldsymbol{\theta}$ 的初始值是：$\boldsymbol{\theta}^{(0)}=(\boldsymbol{\theta}_1^{(0)},\cdots,\boldsymbol{\theta}_d^{(0)})$。

(2) 为了利用从下述 d 个条件分布中所得的 d 个采样生成 $\boldsymbol{\theta}^{(1)}=(\boldsymbol{\theta}_1^{(1)},\cdots,\boldsymbol{\theta}_d^{(1)})$，下面迭代涉及连续不断重新访问 $\boldsymbol{\theta}$ 的所有元素：

$$p(\boldsymbol{\theta}_1^{(1)}|\boldsymbol{\theta}_2^{(0)},\cdots,\boldsymbol{\theta}_d^{(0)})$$
$$p(\boldsymbol{\theta}_2^{(1)}|\boldsymbol{\theta}_1^{(1)}, \boldsymbol{\theta}_3^{(0)},\cdots,\boldsymbol{\theta}_d^{(0)})$$
$$\vdots$$
$$p(\boldsymbol{\theta}_d^{(1)}|\boldsymbol{\theta}_1^{(1)},\boldsymbol{\theta}_2^{(1)},\cdots,\boldsymbol{\theta}_{d-1}^{(1)})$$

(3) 返回步骤 1，重新在 $\boldsymbol{\theta}^{(1)}$ 处初始化向量 $\boldsymbol{\theta}$，并通过步骤 2 进行循环，再次获得新的采样 $\boldsymbol{\theta}^{(2)}$。重复上述步骤，直到收敛为止。

吉尔克斯等人(Gilks et al.，1996，第 7 页)曾提供平稳分布是后验的陈述的概括性证明。在收敛之后，从目标联合后验中就可采样。格曼和格曼(Geman and

---

[1] 又称为取样器。——译者注

Geman，1984)已经证明，随机序列{$\boldsymbol{\theta}^{(n)}$}是具有正确平稳分布的马尔可夫链。盖尔芬德和史密斯(Gelfand and Smith，1990)已经证明，在某些条件下，当来自条件分布所有集合中的循环次数趋于无穷时，链会收敛到平稳后验分布。也可参见坦纳和旺(Tanner and Wong，1987)。一旦出现收敛，可大量采样，并用于计算边缘分布或联合分布的后验矩的样本类似形式。

这里提及的一些结果没有告诉我们，达到收敛需要多少次循环，它是模型相依的。非常重要的是，确保为使链收敛而实施足够的循环次数。可以利用收敛(**convergence**)的各种诊断来进行检验。因为对后验矩进行估计应该建立在从后验分布中获得的采样基础上，标准的做法是抛弃链的较前面的结果，这就是所谓的演练阶段[1](**burn-in phase**)。

对序列模拟算法可加以修改，以使每个采样不直接依赖于紧密接邻的前面采样，却还是依赖于较早采样，一个重要要求是，对后验的当前近似加以改进的概率应是正的且(更可取地)大的。一个更受到限制的马氏性质的吸引人之处是，它会使得对转移分布收敛到目标后验的证明变得容易。

对于贝叶斯分析来说，当联合后验不容易处理，但完全条件分布却是可利用的方便形式时，吉布斯抽样器就可派上用场。许多应用都运用大量技巧与共轭先验的知识及有关的贝叶斯结果，很多都源自较早的预模拟文献，以便设定会产生已知的完全条件分布的先验。

下面考察 MCMC 方法应用的两个例子。

线性回归例子

在 13.3.2 节，我们曾分析给定正态—伽玛先验共轭时，正态线性同方差回归模型的后验分布。可以证明，给定 σ^{-2}时 $\boldsymbol{\beta}$ 的条件后验是多元正态的，而给定 $\boldsymbol{\beta}$ 时 σ^{-2} 的条件后验是伽玛分布的。即使积分是可行的，并且我们能以显性方式推导出后验[参见式(13.32)]，实际上更容易的方法也要使用吉布斯抽样器从联合后验分布中采集大样本。链是由从以精确度参数 σ^{-2} 为条件的正态分布与以 $\boldsymbol{\beta}$为条件的伽玛分布中递推采样构成的。

算法的结构类似于稍后 13.6 节给出的关于两个方程看似不相关回归模型的更为复杂情况。

在许多情况下，当然，以参数分块[2](**blocks**)方式加以研究。例如，在含有非对角线同期协方差矩阵的多个方程变量线性回归模型中，条件均值参数($\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\cdots$)形成一个参数分块，而 $\boldsymbol{\Sigma}$ 形成第二个分块。然后，完全条件分布拥有 $\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\cdots|$数据，$\boldsymbol{\Sigma}$ 以及 $\boldsymbol{\Sigma}$ |数据，$\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\cdots$形式。奇布和格林伯格(Chib and Greenberg，1996，第 418～419 页)对这种情况提供了吉布斯算法的一个纲要。

分层先验例子

在分层先验模型的分析中，吉布斯抽样器取得了很大程度的成功。由式(13.39)至式(13.41)给出的线性分层模型的结构，可以发现，在此情况下，用公式

[1] 又称为预烧，该术语源自工程，意指对某个设备加以调试，为正式运行做的前期准备。——译者注

[2] 又称为分组。——译者注

表示建立在完全条件分布集合上的马尔可夫链是可行的。同样的一般方法能被推广到非线性分层先验模型上，但如果出现非线性以及潜变量模型，就不可避免有另外一些步骤[艾伯特(Albert, 1988)]。

13.5.3 *梅特罗波利斯算法*

吉布斯抽样器是最著名的 MCMC 算法。不过，它的应用性是有限的，因为它要求直接从完全条件分布中采样，而完全条件分布可能不是已知的。允许 MCMC 更一般应用的两个推广是，梅特罗波利斯算法与梅特罗波利斯—黑斯廷斯算法。奇布和格林伯格(Chib and Greenberg, 1995)提供了指导手册与参考文献。假如读者要探索更完整的理解，下面的概述虽然比较简单，却避开必需的许多详细内容。

梅特罗波利斯算法构造一个序列$\{\boldsymbol{\theta}^{(n)}, n=1,2,\cdots\}$，它的分布收敛到目标后验，假定此目标后验是可计算的，并且计算结果至多差一个正规化常值。

为了记号简单，我们再次不用 $p(\boldsymbol{\theta}|\mathbf{y})$ 对 $\mathbf{y}$ 的相依性。此算法由下述步骤构成：

1. 从对 $p(\boldsymbol{\theta}^{(0)})>0$ 的后验最初近似中采样一个起点 $\boldsymbol{\theta}^{(0)}$。例如，采样从边缘后验分布的众数为中心的多元变量 t 分布中得到。

2. 设 $n=1$。从对称的跳跃分布(**jumping distribution**)中采样 $J_1(\boldsymbol{\theta}^{(1)}|\boldsymbol{\theta}^{(0)})$，该分布对于任意序对($\boldsymbol{\theta}^a$, $\boldsymbol{\theta}^b$)具有 $J_n(\boldsymbol{\theta}^a|\boldsymbol{\theta}^b)=J_n(\boldsymbol{\theta}^b|\boldsymbol{\theta}^a)$的性质。一个例子是 $\boldsymbol{\theta}^{(1)}|\boldsymbol{\theta}^{(0)}\sim\mathcal{N}[\boldsymbol{\theta}^{(0)}, \mathbf{V}]$，对于某个固定的 $\mathbf{V}$。跳跃分布的对称性会产生简单性，否则不需要它。

3. 计算密度比值 $r=p(\boldsymbol{\theta}^*)/p(\boldsymbol{\theta}^{(0)})$。

4. 设：

$$\boldsymbol{\theta}^{(1)}=\begin{cases}\boldsymbol{\theta}^*, & \text{以概率 } \min(r,1)\\ \boldsymbol{\theta}^{(0)}, & \text{以概率}(1-\min(r,1))\end{cases}$$

这意味着，采样 $\boldsymbol{\theta}^{(1)}$ 是从具有成分 $\boldsymbol{\theta}^*$ 与 $\boldsymbol{\theta}^0$ 的混合分布中抽到。

5. 回到步骤 2，增大计数器，然后重复下述步骤。

6. 在迭代适当多次数之后，执行分布收敛的必要检查。当收敛出现时，目标后验就会重新获得。

可将该算法看成对 $p(\boldsymbol{\theta})$求最大值的迭代法。如果 $\boldsymbol{\theta}^*$ 使 $p(\boldsymbol{\theta})$增大，那么总是有 $\boldsymbol{\theta}^{(n)}=\boldsymbol{\theta}^*$，然而，如果 $\boldsymbol{\theta}^*$ 使 $p(\boldsymbol{\theta})$减少，那么以概率 $r<1$ 有 $\boldsymbol{\theta}^{(n)}=\boldsymbol{\theta}^*$。

此算法思想，类似于筛选抽样(参见 12.8 节)，尽管这里没有要求：跳跃分布的固定倍数(重数)总要覆盖后验。

梅特罗波利斯算法会生成具有可逆性和不可约性的马尔可夫链，以及确保收敛到平稳分布的哈里斯递归(Harris recurrence)。格尔曼等人(Gelman et al., 1995)已经证明，这个平稳分布是人们所期望的如下后验 $p(\boldsymbol{\theta})$。设 $\boldsymbol{\theta}_a$ 与 $\boldsymbol{\theta}_b$ 是两个点，使得 $p(\boldsymbol{\theta}_b)\geqslant p(\boldsymbol{\theta}_a)$。如果 $\boldsymbol{\theta}^{(n-1)}=\boldsymbol{\theta}_a$ 且 $\boldsymbol{\theta}^*=\boldsymbol{\theta}_b$，那么肯定有 $\boldsymbol{\theta}^{(n)}=\boldsymbol{\theta}_b$，而且 $\Pr[\boldsymbol{\theta}^{(n)}=\boldsymbol{\theta}_b, \boldsymbol{\theta}^{(n-1)}=\boldsymbol{\theta}_a]=J_n(\boldsymbol{\theta}_b|\boldsymbol{\theta}_a)p(\boldsymbol{\theta}_a)$。如果次序被颠倒，同时 $\boldsymbol{\theta}^{(n-1)}=\boldsymbol{\theta}_b$ 且 $\boldsymbol{\theta}^*=\boldsymbol{\theta}_a$，若给定对称跳跃分布的假定，则以概率 $r=p(\boldsymbol{\theta}_a)/p(\boldsymbol{\theta}_b)$，有 $\boldsymbol{\theta}^{(n)}=\boldsymbol{\theta}_a$，并且

$\Pr[\boldsymbol{\theta}^{(n)}=\boldsymbol{\theta}_a, \boldsymbol{\theta}^{(n-1)}=\boldsymbol{\theta}_b]=J_n(\boldsymbol{\theta}_a|\boldsymbol{\theta}_b)p(\boldsymbol{\theta}_b)[p(\boldsymbol{\theta}_a)/p(\boldsymbol{\theta}_b)]=J_n(\boldsymbol{\theta}_a|\boldsymbol{\theta}_b)p(\boldsymbol{\theta}_a)=J_n(\boldsymbol{\theta}_b|\boldsymbol{\theta}_a)p(\boldsymbol{\theta}_a)$。因此，$\boldsymbol{\theta}^{(n)}$ 的边缘分布与 $\boldsymbol{\theta}^{(n-1)}$ 的边缘分布相等，因为它们的联合分布是对称的，所以 $p(\boldsymbol{\theta})$ 是马尔可夫链的对称平稳分布。

13.5.4 *梅特罗波利斯—黑斯廷斯算法*

梅特罗波利斯算法的效果随着对初始近似分布的选择以及对跳跃分布的选择而变化。一个潜在问题是，梅特罗波利斯算法可能会很慢，正如下述情况：通常，当从当前值到一个新值的移动很少发生时，该链变动极小。通过允许使用不是对称的跳跃分布，使算法速度加快。

梅特罗波利斯—黑斯廷斯算法[**Metropolis - Hastings (M - H) algorithm**]与梅特罗波利斯算法一样，区别只是，第 2 步骤中跳跃分布不必是对称的，第 3 步骤对于一般 n 来说，接收概率 r 变成：

$$r_n=\frac{p(\boldsymbol{\theta}^*)/J_n(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(n-1)})}{p(\boldsymbol{\theta}^{(n-1)})/J_n(\boldsymbol{\theta}^{(n-1)}|\boldsymbol{\theta}^*)}=\frac{p(\boldsymbol{\theta}^*)J_n(\boldsymbol{\theta}^{(n-1)}|\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^{(n-1)})J_n(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(n-1)})}$$

其余步骤利用这种改动定义执行。注意到，若任何正规化的常值或者出现在 $p(\cdot)$ 之中，或者出现在 $J_n(\cdot)$ 之中，则在对 r_n 的这种定义中去掉它们。因此，后验概率与跳跃概率仅仅要求计算到该常值为止。参见黑斯廷斯(Hastings, 1970)。

13.5.5 *M - H 例子*

就从后验中获得除了尽可能使用吉布斯抽样器之外，人们期望的采样所需要的采样次数而言，各种不同的跳跃分布会产生具有不同效率的各种不同 M - H 算法。我们给出几个例子，注意到，对于选择跳跃分布来说，存在几个可用的一般性指南。

吉布斯抽样器是 M - H 算法的一种特殊情况。若将 $\boldsymbol{\theta}$ 分割成 d 个分块，则算法第 n 步骤存在 d 个梅特罗波利斯步。跳跃分布是 13.5.2 节给出的条件分布，可以证明，其接收概率总是 1。吉布斯抽样，也称为交错条件抽样(**alternating conditional sampling**)。

借助各种不同转换核用于参数的不同子集上，一种可能方式是使用混合策略。例如，M - H 步骤能与吉布斯抽样器组合起来，后者用于那些可能采用直接抽样的成分。

无关链[1](**independence chain**)是从固定密度 $g(\boldsymbol{\theta})$ 中全部取样，比如说，在接收概率简化成重要权数 $r_n=w(\boldsymbol{\theta}^*)/w(\boldsymbol{\theta}^{(n-1)})$ 比值的情况下。随机游走链[2](**random walk chain**)是令采样 $\boldsymbol{\theta}^*=\boldsymbol{\theta}^{(n-1)}+\varepsilon$，其中，$\varepsilon$ 表示从 $g(\varepsilon)$ 获得的采样。

格尔曼等人(Gelman et al., 1995, 第 334 页)曾经考察对带有方差 $\boldsymbol{\Sigma}$ 的 q 变量正态进行模拟。对于具有跳跃分布 $\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(n-1)}\sim\mathcal{N}[\boldsymbol{\theta}^{(n-1)}, c^2\boldsymbol{\Sigma}]$ 的梅特罗波利斯算法来说，选取 $c\simeq 2.4/\sqrt{q}$，导致了从 q 变量正态进行直接采样时的最大效率。在

[1] 又称为独立链。——译者注

[2] 又称为随机游动链。——译者注

$\boldsymbol{\Sigma}=\sigma^2\mathbf{I}_q$ 情况下，与 $1/q$ 的吉布斯抽样器相比，该效率大约是 0.3。

13.6 MCMC 例子：SUR 吉布斯抽样器

我们阐明吉布斯抽样器应用于看似不相关回归模型的分析。与用于单方程回归相比，这个例子显得更富有挑战性，因为引进了不同方程的相关误差。

考察两个方程的例子，其第 i 个观测值为：

$$y_{1i}=\mathbf{x}_{1i}'\boldsymbol{\beta}_1+\varepsilon_{1i}$$
$$y_{2i}=\mathbf{x}_{2i}'\boldsymbol{\beta}_2+\varepsilon_{2i}$$

其中，$(\varepsilon_1,\varepsilon_2)$表示两变量正态分布，其均值为 0 且协方差矩阵为：

$$\boldsymbol{\Sigma}=\begin{bmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{bmatrix}$$

若对这两个方程合并，则得到其第 i 个观测值：

$$y_i=\mathbf{x}_i'\boldsymbol{\beta}+\varepsilon_i$$

其中，$\varepsilon_i\sim\mathcal{N}[\mathbf{0},\boldsymbol{\Sigma}]$。总之，数据生成过程是：

$$y_i\mid\mathbf{x}_i,\ \boldsymbol{\beta},\ \boldsymbol{\Sigma}\sim\mathcal{N}[\mathbf{x}_i'\boldsymbol{\beta},\ \boldsymbol{\Sigma}]$$

而关注内容在于给定数据 $\mathbf{y}$、$\mathbf{X}$ 时，对回归参数 $\boldsymbol{\beta}$ 与方差参数 $\boldsymbol{\Sigma}$ 的后验均值估计。

考察独立的信息先验，它满足：

$$\boldsymbol{\beta}\sim\mathcal{N}[\boldsymbol{\beta}_0,\ \mathbf{B}_0^{-1}]$$
$$\boldsymbol{\Sigma}^{-1}\sim\text{Wishart}[v_0,\ \mathbf{D}_0]$$

其中，$\mathbf{B}_0$ 被准确定义成先验方差的逆，而由 13.3.5 节定义的逆威沙特则是对逆伽玛的推广。一种可选择的方法，这里没有采用，运用类似于 13.3.2 节那些情况的相依先验，即在设定 ω_0 的 $\boldsymbol{\beta}\mid\boldsymbol{\Sigma}\sim\mathcal{N}[\boldsymbol{\beta}_0,\ \omega_0\boldsymbol{\Sigma}]$情况下。

经过某些代数运算，得到条件后验：

$$\boldsymbol{\beta}\mid\boldsymbol{\Sigma},\ \mathbf{y},\ \mathbf{X}\sim\mathcal{N}\left[\mathbf{C}_0\left(\mathbf{B}_0\boldsymbol{\beta}_0+\sum_{i=1}^{N}\mathbf{x}_i'\Sigma^{-1}\mathbf{y}\right)_i,\mathbf{C}_0\right]$$

$$\boldsymbol{\Sigma}^{-1}\mid\boldsymbol{\beta},\ \mathbf{y},\ \mathbf{X}\sim\text{Wishart}\left[v_0+N,\ \left(\mathbf{D}_0^{-1}+\sum_{i=1}^{N}\mathbf{u}_i'\mathbf{u}_i\right)^{-1}\right]$$

其中，$\mathbf{C}_0=(\mathbf{B}_0+\sum_{i=1}^{N}\mathbf{x}_i'\Sigma^{-1}\mathbf{x}_i)^{-1}$，而 $\mathbf{u}_i=\mathbf{y}_i-\mathbf{x}_i'\boldsymbol{\beta}$。由于条件后验是已知的，同时从两个分布中抽样简单易行，故使用吉布斯抽样器。

就模拟例子而言，我们设每个方程中的回归元都是截距加上单个纯量回归元，这两个方程的回归元是不同的，均生成于标准正态。于是，y_1 与 y_2 是四个回归参数 $\beta_{11}=\beta_{12}=\beta_{21}=\beta_{22}=1$ 生成的，误差方差 $\sigma_{11}=\sigma_{22}=1$，且误差协方差 $\sigma_{12}=\sigma_{21}=-0.5$。样本量或者是 $N=1\,000$，或者是 $N=10\,000$。给定这些数据，我们阐述参数的贝叶斯估计，其中，先验分布设 $\boldsymbol{\beta}_0=\mathbf{0}$，$\mathbf{B}_0^{-1}=\tau\mathbf{I}$，$\mathbf{D}_0=\mathbf{I}$ 而 $v_0=5$。为检查各

种不同先验的影响，考察 τ 的三个值，即 $\tau=10$，$\tau=1$ 以及 $\tau=1/10$，较少的 τ 值对应于较紧密的先验。

吉布斯抽样器递推地从条件后验分布中采样。我们拒绝构成“演练阶段”的前5 000 个复制，然后报告利用后面 50 000 个与 100 000 个复制的结果。

表 13.3 给出节选部分的结果，报告 5 个不同样本中每一个系数边缘后验分布的均值及方差，而 5 个不同样本自身均是独立采样的。前三列阐述各种不同 τ 值的敏感性分析，这表明结果不是非常敏感的。与第一列相比，第四列表明，加倍复制具有非常小的效果。与第一列相比，第五列表明，将样本量增加 10 倍至 10 000[1]会极大提高精确度，如人们所料，将系数的标准误差减少到大于 3 的因子数，仅仅对点估计的影响相对小些。

表 13.3　吉布斯抽样：看似不相关回归例子[a]

先验参数 τ	$\tau=10$	$\tau=1$	$\tau=1/10$	$\tau=10$	$\tau=10$
样本量 N	1 000	1 000	1 000	1 000	10 000
吉布斯样本复制	50 000	50 000	50 000	100 000	100 000
β_{11}（方程 1 的截距）	0.971	1.013	0.983	1.020	1.010
	(0.031 0)	(0.031 2)	(0.031 6)	(0.032 4)	(0.010 0)
β_{12}（方程 1 的斜率）	1.026	0.983 5	1.006	1.006	1.015
	(0.026 5)	(0.027 1)	(0.026 5)	(0.026 8)	(0.008 6)
β_{21}（方程 2 的截距）	1.016	0.972	0.993	1.017	0.991
	(0.030 9)	(0.032 5)	(0.032 2)	(0.032 6)	(0.010 0)
β_{22}（方程 2 的斜率）	0.983	0.992	0.979	1.005	1.007
	(0.025 6)	(0.028 5)	(0.027 2)	(0.027 7)	(0.008 5)
σ_{11}（方程 1 的方差）	0.960	0.969	1.012	1.043	1.010
	(0.042 9)	(0.043 4)	(0.045 3)	(0.046 6)	(0.014 3)
σ_{12}（误差协方差）	−0.499	−0.507	−0.576	−0.576	−0.515
	(0.034 0)	(0.035 8)	(0.036 8)	(0.037 9)	(0.011 3)
σ_{22}（方程 1 的截距）	0.950	1.066	1.049	1.062	1.002
	(0.425)	(0.047 6)	(0.046 7)	(0.047 2)	(0.014 1)

[a] 模型是两个看似不相关回归的方程。该表给出了每个参数后验分布的均值与标准差。较小的 τ 值对应较紧凑的先验。

一种检查收敛方式是考察输出的均值与标准差，看看它们是否漂动或停留在同一水平上。当变动很小，比如说就 10 000 复制而言小于 0.1，则认为出现收敛。人们也可同时考察几个链。这些采样总是相关的，但一个重要问题是，自相关函数会怎样快地衰落至 0。有时，此问题不是固定的，而且它自然是算法所固有的。人们还能采用每 1/10 或每 1/100 观测值来消除序列相关。

为检查吉布斯抽样器是否收敛到目前情况下的平稳后验分布上，我们计算从每个系数收敛后的后验中获得的采样之自相关系数的前 20 个。缺乏收敛会通过从目标分布中得到的采样存在序列相关而显示出来。当复制次数很少时，比如说

〔1〕 原著中这里为 100 000，但应为 10 000。——译者注

1 000，在一些情况下，可以发现，自相关系数高达 0.06。不过，当复制次数为50 000或更大时，实际上没有直到 20 阶的序列相关证据，而相关性会随阶数而消失。在大多数情况下，估计值比 0.005 更小一些。容易验证，对于 $N=1\,000$，先验系数 τ 对后验具有相当小的影响。这种计算非常简单，使用时仅需花费几秒钟而已。

13.7 数据增广

有时，吉布斯抽样器能用于通过引入辅助变量而得到的更广泛模型上。特别地，这是涉及潜变量的模型情况，诸如离散选择模型、截取与删失模型，以及后面几章将引入的有限混合模型。

在纯量情况下，潜因变量 y^* 是不可观测的；相反，我们仅仅观测到关于某个设定函数 y 的 $y=g(y^*)$。例如，在 logit 或 probit 模型中(参见第 14 章)，仅仅可能观测到 y^* 是正的或负的，在此情况下，$y=\mathbf{1}(y^*>0)$，并且当 $y^*>0$ 时，观测到 $y=1$，而当 $y^*\leqslant 0$ 时，观测到 $y=0$。

潜变量的贝叶斯分析，特别是吉布斯抽样器的应用，均通过用估算值[1] (**imputed values**)代替潜变量而得以实施。倘若我们能依照观测到的变量写出潜变量的预测密度，这一步就可行。添加估算值就好像它们是观测到数据的方法称为数据增广(**data augmentation**)。(一个例子是由 10.3.7 节给出的，其中解释了 EM 算法。)一种深刻观点归功于坦纳和旺(Tanner and Wong, 1987)，即仅仅建立在已观测到数据基础上的后验是难以处理的，但在数据增广之后所得到的后验，若利用吉布斯抽样器，这就常常容易处理。

考察既依据直接观测到的变量 $\mathbf{y}$，又依据潜变量 $\mathbf{y}^*$ 所表述的后验：

$$p(\boldsymbol{\theta}\mid\mathbf{y})=\int_{\mathbf{y}^*}p(\boldsymbol{\theta}\mid\mathbf{y},\mathbf{y}^*)f(\mathbf{y}^*\mid\mathbf{y})d\mathbf{y}^* \tag{13.49}$$

其中，右边积分可被解释成关于 $\mathbf{y}^*$ 的平均运算。

类似于 EM 方法，数据增广涉及在估算步骤(**imputation step**)即 I 步骤与后验步骤(**posterior step**)即 P 步骤之间的循环。

在估算步骤，从 $\mathbf{y}^*$ 的完全条件密度中采样。这是对出现在概率分布中的参数 ψ 加以平均，此概率分布联系 $\mathbf{y}^*$ 与 $\mathbf{y}$。其预测分布是：

$$f(\mathbf{y}^*\mid\mathbf{y})=\int_{\psi}f(\mathbf{y}^*\mid\mathbf{y},\psi)f(\psi\mid\mathbf{y})d\psi \tag{13.50}$$

给定当前来自 $p(\boldsymbol{\theta}\mid\mathbf{y})$ 的采样，能从 $f(\mathbf{y}^*\mid\mathbf{y})$ 中得到采样，为了获得 m 重新估算 $\mathbf{y}_i^*$，$i=1,\cdots,m$，就要重复该步骤 m 次。这就完成 I 步骤。

给定来自 I 步骤的数据增广，P 步骤通过对 $p(\boldsymbol{\theta}\mid\mathbf{y})$ 的当前近似更新来实施；因而有：

---

[1] 又称为借补值。——译者注

$$更新\ p(\boldsymbol{\theta}|\mathbf{y})=\frac{1}{m}\sum_{i=1}^{m}p(\boldsymbol{\theta}|\mathbf{y},\mathbf{y}_i^*) \tag{13.51}$$

然后，算法返回到 I 步骤。

当 $m=1$ 时，此方法相当于通过吉布斯抽样实施积分式(13.49)。若选取 m 充分大，则后验分布就近似得更好些。把数据增广方法应用于缺失数据问题的扩展例子，将在第 26 章给出。

13.8 贝叶斯模型选择

第 7 章与第 8 章已经研究了假设检验、设定诊断以及与源自频率学派观点的模型比较问题。在本节，我们考察最重要的工具——贝叶斯因子(**Bayes factors**)，运用它表示贝叶斯分析计算支持零假设(模型)证据的效力。它还可作为模型选择的准则，而不管所考虑的内容是嵌入式模型还是非嵌入式模型对。在经济计量学文献中，泽尔纳(Zellner, 1971, 1978)已经提供了模型选择内容的早期讨论。我们这里的研究是建立在卡斯和拉夫特里(Kass and Raftery, 1995)的评述性论文基础上。

用 $\mathbf{y}$ 表示数据，而所考虑的两个假设可能是非嵌套的，分别用 H_1 与 H_2 表示。两个假设的先验概率是 $\Pr[H_1]$ 与 $\Pr[H_2]$。相对应的数据生成过程是 $\Pr[\mathbf{y}|H_1]$ 与 $\Pr[\mathbf{y}|H_2]=1-\Pr[\mathbf{y}|H_1]$。模型的先验概率，通过反映在似然中样本证据转换成后验概率。由贝叶斯定理知：

$$\Pr[H_k|\mathbf{y}]=\frac{\Pr[\mathbf{y}|H_k]\Pr[H_k]}{\Pr[\mathbf{y}|H_1]\Pr[H_1]+\Pr[\mathbf{y}|H_2]\Pr[H_2]},\quad k=1,2 \tag{13.52}$$

以及后验优势比(**posterior odds ratio**)：

$$\frac{\Pr[H_1|\mathbf{y}]}{\Pr[H_2|\mathbf{y}]}=\frac{\Pr[\mathbf{y}|H_1]\Pr[H_1]}{\Pr[\mathbf{y}|H_2]\Pr[H_2]}\equiv B_{12}\frac{\Pr[H_1]}{\Pr[H_2]} \tag{13.53}$$

其中，$B_{12}=\Pr[\mathbf{y}|H_1]/\Pr[\mathbf{y}|H_2]$，称为贝叶斯因子。当后验优势比大于 1 时，假设 1 就更可取。式(13.53)的右边将后验优势比表述成贝叶斯因子与先验优势比的乘积。如果两个先验模型相等，因而 $\Pr[H_1]=\Pr[H_2]$，贝叶斯因子等于后验优势比，这支持了 H_1。若涉及几个假设，则贝叶斯因子用于这些假设的所有序对的计算。即使假设不是嵌套的，也可定义贝叶斯因子。

贝叶斯因子具有似然比形式。它依赖于未知参数，用向量 $\boldsymbol{\theta}_1$ 与 $\boldsymbol{\theta}_2$ 表示未知参数，它们可通过在参数空间上关于先验进行平均或积分加以剔除，所以：

$$\Pr[\mathbf{y}|H_k]=\int\Pr[\mathbf{y}|\boldsymbol{\theta}_k,H_k]\pi(\boldsymbol{\theta}_k|H_k)d\boldsymbol{\theta},\quad k=1,2 \tag{13.54}$$

由 13.2.5 节知，式(13.54)提供了给定先验分布时数据的边缘与预测概率。

一种新的困难是，这个表达式依赖于出现在似然中的所有常值。当计算后验时，可忽略这些常值。但是，计算贝叶斯因子时却需要它们。如果式(13.54)积分没有可利用的显性解，例如，重要抽样，就要求运用数值计算。卡斯和拉夫特里

(Kass and Raftery, 1995)曾经评论到,贝叶斯因子计算方面存在大量文献,这里我们将不去继续讨论这一内容。注意到,存在某些对贝叶斯因子的渐近近似,是很容易利用来自对似然求极大值的软件输出加以计算。

按照反对 H_2[1] 的证据,可对贝叶斯因子给出解释,"将此与另一种理论相对比,贝叶斯因子是由数据提供支持、由统计模型表述的一种科学理论的证据汇总"[卡斯和拉夫特里(Kass and Raftery,1995,第 777 页)]。在频率学派分析中,两倍对数似然比是经常使用的量。类似地,两倍的贝叶斯因子对数作为计算证据时所使用的准则。卡斯和拉夫特里阐述了,反对零假设证据效力的下述分类,这是在他们自己研究时确立的有用工具,参见表 13.4。

表 13.4 对贝叶斯因子的解释

贝叶斯因子	$2\ln(B_{12})$	对应于 H_1 的证据
1～3	0～2	弱
3～20	2～6	正
20～150	6～10	强
>150	>10	非常强

假设正在比较的两个模型是嵌套的。用 H_0 表示约束模型 H,而用 H_1 表示无约束模型。利用后验优势比对两个模型进行成对比较,正如前面所证明的,需要计算贝叶斯因子。对零假设模型的贝叶斯因子可定义成:

$$B_{01}=\frac{m(\mathbf{y}|H_0)}{m(\mathbf{y}|H_1)}$$

其中,$m(\mathbf{y}|H_j)$表示模型设定 H_j 的边缘似然。若模型 H_0 与 H_1 都是嵌套的,则采用 Savage-Dickey 密度比方法[参见威迪内里和沃瑟曼(Verdinelli and Wasserman, 1995)]计算此贝叶斯因子。

不管模型是嵌套的,还是非嵌套的,由奇布(Chib, 1995)提出的重要思想用于计算贝叶斯因子,比早期文献所建议的计算方法更为简便。他的方法是由两种有关思想构成的。对于给定的模型 H_k 来说,首先将边缘密度 $m(\mathbf{y})$重新写成一个比值:

$$m(\mathbf{y})=\frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta}|\mathbf{y})}$$

其中,分子是密度(包括常值)与先验的乘积,而分母是 $\boldsymbol{\theta}$ 的后验密度。这个结果是式(13.1)中项的重新,限制条件是我们使用记号 $m(\mathbf{y})$代替 $f(\mathbf{y})$或较早使用的 $\Pr[\mathbf{y}|H_k]$;它仅仅表明,边缘密度是一个正规化的常值。其次,在成功应用 MCMC 算法之后,我们将在给定点 $\tilde{\boldsymbol{\theta}}$ 上利用后验密度估计 $\pi(\tilde{\boldsymbol{\theta}}|\mathbf{y})$的蒙特卡罗估计值。由此可得:

$$\ln \hat{m}(\mathbf{y})=\ln f(\mathbf{y}|\tilde{\boldsymbol{\theta}})+\ln \pi(\tilde{\boldsymbol{\theta}})-\ln \pi(\tilde{\boldsymbol{\theta}}|\mathbf{y}) \tag{13.55}$$

[1] 原著中此处为 H_1,应为 H_2。——译者注

因此，给定右边一些项的估计值，边缘密度能很容易地利用来自吉布斯抽样器输出加以计算。然而，该方法被奇布和叶利阿泽科夫（Chib and Jeliazkon, 2001）推广到输出是由梅特波罗利斯—黑斯廷斯算法给出的情况。

在复杂且高度参数化的模型中，对贝叶斯因子计算是一件不简单的事。不过，可以证明，施瓦茨准则也是著名的贝叶斯信息准则（参见 8.5 节），它会给出对贝叶斯因子对数的大致近似。回顾，$\text{BIC}=-2\ln L(\hat{\boldsymbol{\theta}}_{\text{ML}})+\ln Nq$。当可以利用对数似然值时，这很容易计算出来。

由式（13.52）知，很明显，模型的先验概率比在计算反对零假设证据中起作用。在许多情况下，研究者很少会继续指派这些概率。此种考虑在研究贝叶斯因子对先验模型概率敏感性的文献中受到某种关注。

13.9 应用研究

在贝叶斯文献中，马尔可夫链运用现今已成为主流。因为该方法是密集计算，好的软件包是基础性的。在写作成书时，WinBUGS 软件包，即 BUGS（利用吉布斯抽样进行贝叶斯推断）的最新版本，受到广泛推荐，而且发现，它对分层模型和缺失数问题特别有用。在 BUGS 网站上就可以利用它。有关其他贝叶斯软件包的更详细信息，参见盖默曼（Gamerman, 1997, 5.6 节）。

将多长的连续不断的马尔可夫链用于执行的问题是研究中的一个活跃领域。需要提及的是，一些诊断检查可用于判断是否收敛，但是它们常常不具有普适的可应用性。卡佩和罗伯特（Cappè and Robert, 2000）曾提供了包括停止规则的实施问题的一个回顾。显然，条件分布的复杂性是一个重要因素。源自马尔可夫的纯量参数输出图形是证实收敛的可视化吸引人的方法，但可利用一些更正式的方法[格韦克（Geweke, 1992）]。另一个由格尔曼和鲁宾（Gelman and Rubin, 1992）给出的建议是使用多重（平行的）吉布斯抽样器，每一个都可从不同的初始值开始，看看各种不同的链是否收敛到同样的后验分布。泽尔纳和敏（Zellner and Min, 1995）曾提出几种收敛准则，若后验分布能以显性方式写出，就运用它们。

13.10 文献注释

有几部优秀的长篇论著强调了贝叶斯分析现代计算方法，这些著作包括盖默曼（Gamerman, 1997）与格尔曼等人（Gelman et al., 1995）的书。相对容易入门的研究著作是，吉尔（Gill, 2002）、库普（Koop, 2003）、兰开斯特（Lancaster, 2004）的书。库普曾阐述许多标准非线性横截面模型与面板数据的一些贝叶斯方法。而泽尔纳（Zellner, 1971）与利默（Leamer, 1978）撰写的书仍是有价值结果的来源。

13.2 斯蒂格勒（Stigler, 1986）提供了贝叶斯（Bayes, 1764）研究工作的良好解释。贝叶斯第一次阐述了概率的某些性质，即著名的 $\Pr[A\mid B]=\Pr[A\cap B]/\Pr[B]$。然后，贝叶斯利用这一结果来获得后验概率 $\Pr[a<\theta<b\mid y]$，其中，a 与 b 被设定为有界的，y 表示 N 个二项试验的成功次数，而 θ 表示每次成功的未知概

率。贝叶斯选择均匀先验,在此情况下,后验密度 $f(\theta|y)\propto f(y|\theta)$。贝叶斯的例子是富于挑战性的,因为它没能准确计算后验概率,它涉及不完全伽玛,直到 20 世纪才把它列成表。最初,贝叶斯论文被人们忽略了。归功于拉普拉斯和其他学者的更为广泛使用的方法是逆概率方法,即设 $f(\theta|y)\propto f(y|\theta)$。这些方法可由极大似然法来代替,极大似然法由费希尔(Fisher, 1922)引进,他的论文直接批评了贝叶斯方法及逆概率方法。

海德和约翰斯通(Heyde and Johnstone, 1979)已经讨论有关收敛到后验正态性的正则条件。特雷恩(Train, 2003)提供了所谓的贝伦斯坦—冯 · 米泽斯定理的优秀但稍欠正式的处理。

13.3 泽尔纳(Zellner, 1971)与利默(Leamer, 1978)均是线性回归贝叶斯分析的优秀来源。

13.4 格韦克(Geweke, 1989)与格韦克和基恩(Geweke and Keane, 2001)均是关于蒙特卡罗积分的珍贵的参考文献。

13.5 卡塞拉和乔治(Casella and George, 1992)曾经提供吉布斯抽样器的解释性处理。由奇布及其合作者以及格韦克及其合作者撰写的大量论文,涵盖了微观经济计量学中许多有意思的专题。奇布和格林伯格(Chib and Greenberg, 1996, 第 3 节)曾提供 MCMC 的一系列应用,包括看似不相关回归模型以及 Tobit 模型和 probit 模型。在后者情况下,他们证明了由于把吉布斯抽样与数据增广结合起来的方法而引起的计算简化。数据增广可用于处理为了研究许多删失模型与离散选择模型中自然出现的基本不可观测变量而引入的潜变量问题。奇布(Chib, 2001)提供了包括许多导致线性与非线性模型的 MCMC 算法的详细而最新的综述。格韦克和基恩(Geweke and Keane, 2000)专门研究了积分方法;其内容既涵盖贝叶斯专题,又涵盖非贝叶斯专题。

习 题

13-1 证明,如果 $\boldsymbol{\beta}|\lambda\sim\mathcal{N}[\boldsymbol{\mu}, \lambda^{-1}\boldsymbol{\Sigma}]$,同时 $\lambda\sim\text{Gamma}[\alpha/2, \alpha/2]$,那么 $\boldsymbol{\beta}$ 的无条件分布是具有参数$(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha)$的多元变量 t 分布。

13-2 [源自奇布(Chib, 1992)。]考察删失回归或 Tobit 模型(参见 16.3 节),其中,$y^*=\mathbf{x}'\boldsymbol{\beta}+\varepsilon$, $\varepsilon\sim\text{iid }\mathcal{N}[0, \sigma^2]$,而且当 $y^*>0$ 时,y 是可观测的,当 $y^*\leqslant 0$ 时,y 是不可观测的(删失的)。关于 y,存在 N_0 个删失可观测值,并用 y_0 意指它们。引入对应于删失观测值的潜变量 z,使得如果第 i 个观测值属于删失集合,则 $z_i<0$。数据增广方法可用于推导潜变量 $-\infty<z_i<0$,独立随机变量的集合分布作为截取正态分布,其支集为$(-\infty,0)$,而 pdf $\phi(z_i|y_i,\boldsymbol{\beta},\sigma^2)/(1-\Phi(\mathbf{x}_i'\boldsymbol{\beta}/\sigma))$,$-\infty<z_i<0$,其中,$\phi$ 与 Φ 分别为正态的 pdf 与 cdf。运用 $\boldsymbol{\beta}$ 的正态先验以及 σ^{-2} 的伽玛先验。

(a) 证明,设定关于 z_i、$\boldsymbol{\beta}$ 以及 σ^{-2} 的完全条件集合是可行的。

(b) 运用 (a)部分的结果,概述模拟 z_i、$\boldsymbol{\beta}$ 以及 σ^{-2} 的吉布斯算法。

(c) 解释如何获得 $\boldsymbol{\beta}$ 与 σ^{-2} 的合适初始值。

第四部分 横截面数据模型

第四部分由第 14 章～第 20 章构成，内容涵盖横截面数据的核心内容，即非线性受限因变量模型，这类模型是通过因变量取值范围来定义。涉及的专题包括：二值数据、多项式数据、持续期限数据和计数数据，以及对删除、截取以及样本选择的复杂情况的研究。第四部分的核心基础是，最小二乘法与极大似然估计。

第 14 章和第 15 章涵盖二值数据与多项式数据，它们是离散结果及离散选择分析中的标准形式。极大似然方法占据主导地位。在这些模型中，对条件概率进行各种参数化会产生各类不同模型，譬如著名的 logit 模型与 probit 模型，这两个是得到公认的。最近文献关注含有更灵活的条件概率函数形式的约束较少的建模，并且可并入不可观测异质性。这些目标激励了半参数方法以及基于模拟估计方法的运用。

删失、截取或者样本选择组成了第 16 章分析的几种重要模型类型。建立很久的 Tobit 模型是这方面文献的核心，但它的估计及推断却依赖于强分布假设，以便获得一致估计。我们还考察一些较为新颖的半参数方法，这些方法依赖于较弱假设。

第 17 章～第 19 章考察持续期限模型，关注内容既有时期长度的确定，比如失业时期的长短，又有对从一个初始状态转移到另一个状态风险率的建模。这种分析，既包括离散时间模型又包括连续时间模型，既有参数公式又有半参数公式，包括一些标准模型像指数、威布尔模型以及比例风险模型。第 18 章涵盖了不可观测异质性的大量模型公式及解释。状态相依性和不可观测异质性作为时期平均长度的决定因素，这个方面的相对重要性是一个中心问题，对其求解会产生有关可供选择建模方法的基本问题。第 19 章运用竞争风险公式与多重时期模型，讨论了几种事件类型模型。

第 20 章涵盖健康经济学中非常普遍的事件计数类型分析。在计数数据模型与持续期限模型之间，存在众多紧密联系和平行关系，因为它们在随机过程中拥有共同基础。我们分析了广泛运用的泊松与负二项式回归模型以及一些重要变形，诸如两部分或围栏模型、零膨胀模型、潜类型模型和内生回归元模型，所有这些模型均迎合事件过程的各种不同方面。

14 二值结果模型

14.1 引　论

离散结果(**discrete outcome**)或定性响应模型(**qualitative response models**)是指因变量即关注结果落入 m 个互不相交类型之一的模型。通常不存在关于分类的一个自然排序。例如,对工人职业进行分类化处理。

本章考察最简单的二值结果(**binary outcomes**)情况,其中存在两种可能结果。一些例子包括,一个人是否就业,消费者是否购买。对二值结果进行建模非常简单,而且估计通常利用极大似然法,因为数据分布必须由贝努利模型来定义。如果一个结果的概率等于 p,那么另一个结果的概率必是$(1-p)$。对于一些回归应用来说,概率 p 将随不同个体而变化,作为回归元的函数。两个标准的二值结果模型,即 logit 模型与 probit 模型,设定此概率的不同的函数形式作为回归元的函数。这两个估计量之间的差异在性质上类似,即在最小二乘回归中使用不同函数形式的条件均值。

14.2 节给出一个数据例子。14.3 节对标准模型包括 logit 与 probit 模型的统计结论做一个概述。14.4 节阐述起因于基本潜变量的二值结果模型。将上述内容轻而易举地推广到多项式模型(参见第 15 章)以及关于删失或选取样本的模型(参见第 16 章)时,这一公式极为有用。14.5 节详细阐述,当结果之一被故意过度抽样时,对标准估计方法进行必要的修正。加总问题则在 14.6 节考察。14.7 节讨论对概率 p 模型施加很少结构的二值结果模型的半参数方法。

14.2 二值结果例子:钓鱼方式的选择

本节对租船钓鱼与码头钓鱼之间做选择加以建模。其因变量是一个二值变量,满足:

$$y_i=\begin{cases}1, & 若租船钓鱼\\0, & 若码头钓鱼\end{cases}$$

其中,为了简单起见,选取值为 1 与 0。单个解释变量是 $x_i=\ln \mathrm{relp}_i=\ln(\mathrm{relp}_i)$,这

里，relp 表示租船钓鱼价格与码头钓鱼价格的比值，因而：

$$x_i = \ln \text{relp}_i = \ln(\text{price}_{\text{租船},i}/\text{price}_{\text{码头},i})$$

租船钓鱼与码头钓鱼的价格都会因各种因素导致随不同个体而变化，例如，学钓鱼的起点千差万别。可以认为，租船钓鱼概率将随其相对价格提高而减少。

各种数据已由表 14.1 概括。630 名个体样本是 15.2 节中以较详细方式表述数据的子集，那里考察四种不同的钓鱼方式以及另外的一些回归元。样本的 71.7%个体选择租船钓鱼。对于选取租船钓鱼的人来说，平均而言，租船钓鱼费用小于码头钓鱼费用，因为 75 美元<121 美元。对于选取码头钓鱼的人来说，费用正好相反。所以，看起来价格具有预期效应。

表 14.1　钓鱼方式选择：数据概述

变量	子样本平均值		所有 y
	$y=1$ 租船	$y=0$ 码头	
租船价格(美元)	75	110	85
码头价格(美元)	121	31	95
ln relp	−0.264	1.643	0.275
样本概率	0.717	0.283	1.000
观测值	452	178	630

y_i 对 x_i 的 OLS 回归(**OLS regression**)忽略因变量的离散性，并没有把预测概率限制在 0 与 1 之间。

一种更合适的模型是 logit 模型(**logit model**)(参见 14.3.4 节)，它设定：

$$p_i = \Pr[y_i = 1 \mid x_i] = \frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)}$$

很明显，这就确保了 $0 < p_i < 1$。运用极大似然估计(参见 14.3.3 节)得到参数估计值，这已由表 14.2 中第一列给出。该 logit 模型蕴含的边际效应等于：

$$\frac{dp_i}{dx_i} = \frac{\exp(\beta_1 + \beta_2 x_i)}{(1 + \exp(\beta_1 + \beta_2 x_i))^2}\beta_2$$

表 14.2　钓鱼方式选择：logit 估计值与 probit 估计值[a]

回归元	logit	probit 模型	OLS
常值	2.053	1.194	0.784
	(12.15)	(13.34)	(65.58)
ln relp	−1.823	−1.056	−0.243
	(−12.61)	(−13.87)	(−28.15)
$-\ln L$	−206.83	−204.41	—
伪 R^2	0.449	0.455	0.463

[a] 若租船钓鱼，则因变量 $y=1$；若码头钓鱼，则 $y=0$。回归元 $x=\ln$ repl 表示租船钓鱼价格相对于码头钓鱼价格的自然对数。括号中含有 t 统计量的截距与斜率参数估计均来自 logit 与 probit 模型的 ML 估计，以及源自 OLS 估计。

由于 $\hat{\beta}_{2,\text{LOGIT}}<0$，正如人们所料，可得 $dp_i/dx_i<0$。边际效应的真实数量会随计算点 x_i 不同而变化（参见 14.3.2 节）。尽管没有涉及其他一些模型，但对 logit 模型的近似是 $dp_i/dx_i \simeq \bar{y}(1-\bar{y})\hat{\beta}_2=-0.370$。不过，OLS 回归却给出了直接估计值 -0.243。

一种可供选择的模型是 probit 模型（**probit model**）（参见 14.3.5 节），该模型设定：

$$p_i=\Pr[y_i=1|x_i]=\Phi(\beta_1+\beta_2 x_i)$$

其中，$\Phi(\cdot)$表示标准正态累积分布函数，因此 $p_i=\int_{-\infty}^{\beta_1+\beta_2 x_i}(2\pi)^{-1/2}e^{-z^2/2}dz$。ML 系数已由表 14.2 中的第 2 列给出，而且显著地不同于 logit 系数。由于不同设定被用于估计之中，故其系数不可对比。这类似于我们不能对具有条件均值 $\mathbf{x}'\boldsymbol{\beta}$ 的模型与具有条件均值 $\exp(\mathbf{x}'\boldsymbol{\beta})$ 的模型进行比较一样。对于 probit 模型来说，$dp_i/dx_i=\phi(\beta_1+\beta_2 x_i)\beta_2$，其中，$\phi(\cdot)$表示标准正态密度。由于 $\hat{\beta}_{2,\text{PROBIT}}<0$，所以再次得出，$dp_i/dx_i<0$。

虽然对于不同模型来说，斜率系数一定会各不一样，可是由表 14.2 知，统计量是相似的，且都是相当大的。probit 模型的对数似然是 2.42，大于 logit 对数似然，由于两个模型使用相同的参数值，所以这支持了 probit 模型。在许多其他例子中，就不同模型来说，$\ln L$ 的差异会很小。可将源自三个模型的预测概率作为 x 的函数，画在图 14.1 中。对于 OLS，我们假定 $\Pr[y_i=1|x_i]=\beta_1+\beta_2 x_i$ 关于 x_i 是线性的，而 logit 与 probit 的非线性函数基本上是等价的。

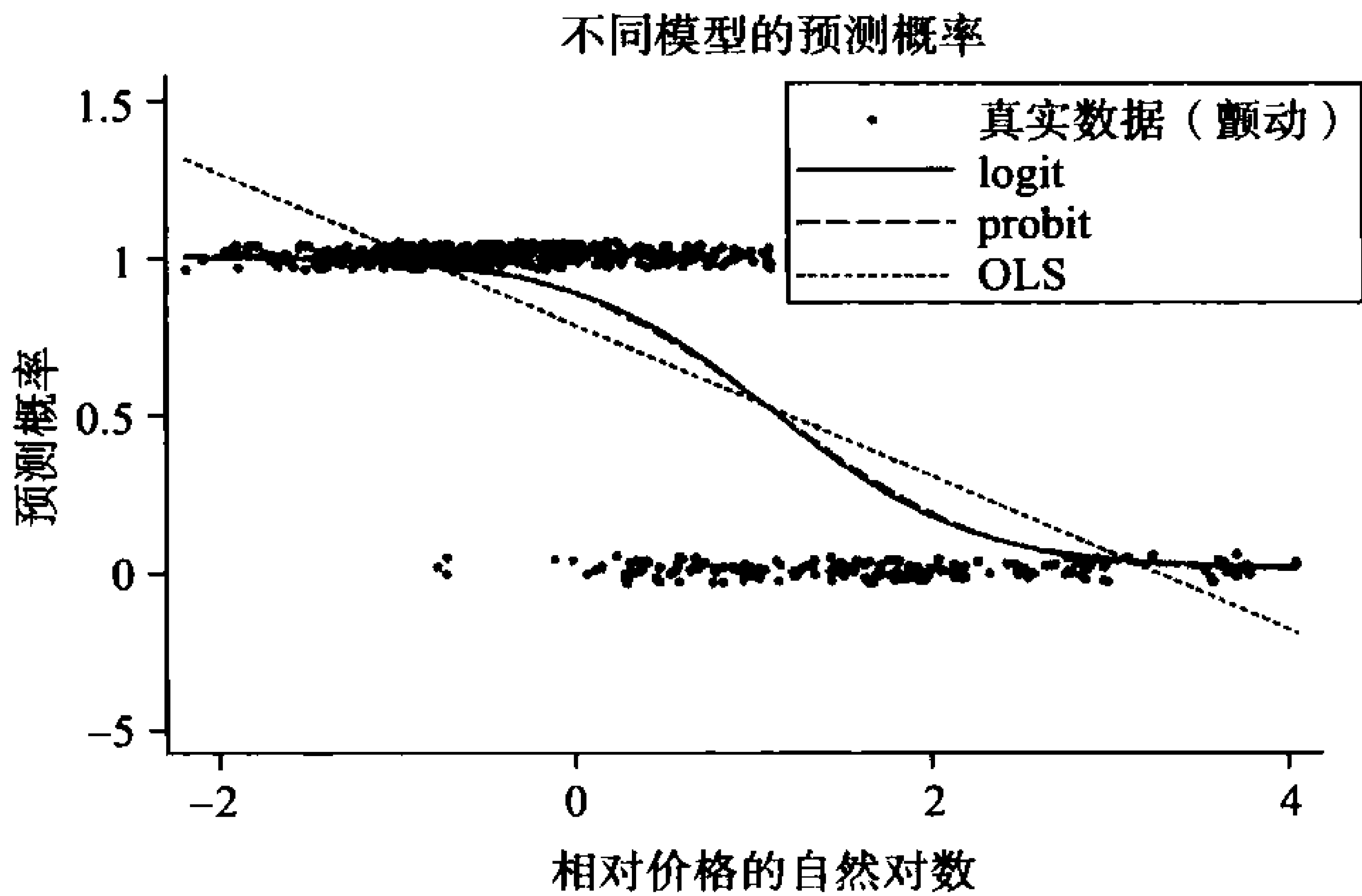


图 14.1 租船钓鱼：当单个回归元是相对价格自然对数时，来自 logit 与 probit 模型的预测概率以及 OLS 预测。为了可读性，在颤动之后，画出 1 或 0 的实际结果。数据由 620 位个体组成。

14.3 logit 模型与 probit 模型

现在给出这些模型的更为正式的理论。我们将统计学引论中掷硬币的二值结果，直接推广到将成功概率建模成依赖回归元的情况。两种普遍运用的参数化方法都会产生 logit 与 probit 模型。若利用潜变量，则关于这些参数化的动机推迟到

14.4 节加以阐述。

14.3.1 一般二值结果模型

对于二值结果来说，因变量 y 取两个值之一。我们设：

$$y=\begin{cases}1, & \text{以概率 } p\\ 0, & \text{以概率 } 1-p\end{cases}$$

为了不失一般性，在设置值为 1 与 0 的背景下，需要建模的内容是 p，即决定结果的概率。在统计学引论中，该模型被表述成掷硬币的结果，其中，正面向上导致 $y=1$ 且以概率 p 发生。

回归模型通过对概率 p 进行参数化，使其依赖于回归元 $\mathbf{x}$ 和 $K\times 1$ 维参数向量 $\boldsymbol{\beta}$ 而得以建立。普遍使用的模型是具有条件概率(**conditional probability**)的单指标形式，条件概率由

$$p_i\equiv \Pr[y_i=1\mid \mathbf{x}]=F(\mathbf{x}_i'\boldsymbol{\beta}) \tag{14.1}$$

给出，其中，$F(\cdot)$表示设定函数。为了确保 $0\leqslant p\leqslant 1$，将 $F(\cdot)$设定成累积分布函数是很自然的。

表 14.3 给出最普遍使用的二值结果模型。当 $F(\cdot)$表示逻辑斯蒂分布时，得出 logit 模型[1](**logit model**)，而当 $F(\cdot)$表示标准正态累积分布函数时，得到 probit 模型[2](**probit model**)。注意到，如果 $F(\cdot)$是 cdf，那么这个 cdf 仅仅用于对参数 p 进行建模，且不表示 y 自身的 cdf。当 $F(\cdot)$表示极值分布的 cdf 时，就产生极少运用的互补双对数回归模型(**complementary log-log model**)。它不同于其他一些模型，因为它关于 0 是非对称的，同时当结果之一极少发生时才会使用它。线性概率模型(**linear probability model**)不使用 cdf，反而设定 $p_i=\mathbf{x}_i'\boldsymbol{\beta}$。

表 14.3　二值结果数据：一些常用模型

模型	概率($p=\Pr[y=1\mid\mathbf{x}]$)	边际效应($\partial p/\partial x_j$)
logit	$\Lambda(\mathbf{x}'\boldsymbol{\beta})=\frac{e^{\mathbf{x}'\boldsymbol{\beta}}}{1+e^{\mathbf{x}'\boldsymbol{\beta}}}$	$\Lambda(\mathbf{x}'\boldsymbol{\beta})[1-\Lambda(\mathbf{x}'\boldsymbol{\beta})]\beta_j$
probit	$\Phi(\mathbf{x}'\boldsymbol{\beta})=\int_{-\infty}^{\mathbf{x}'\boldsymbol{\beta}}\phi(z)dz$	$\phi(\mathbf{x}'\boldsymbol{\beta})\beta_j$
互补双对数回归	$C(\mathbf{x}'\boldsymbol{\beta})=1-\exp(-\exp(\mathbf{x}'\boldsymbol{\beta}))$	$\exp(-\exp(\mathbf{x}'\boldsymbol{\beta}))\exp(\mathbf{x}'\boldsymbol{\beta})\beta_j$
线性概率	$\mathbf{x}'\boldsymbol{\beta}$	β_j

14.3.2 边际效应

关注内容是回归元的变化对 $y=1$ 的条件概率的边际效应(**marginal effect**)，对于一般概率模型(14.1)来说，假定第 j 个回归元变化是连续的，得到：

---

[1] 又称为对数单位模型。——译者注
[2] 又称为概率单位模型。——译者注

$$\frac{\partial \Pr[y_i=1\mid \mathbf{x}_i]}{\partial x_{ij}}=F'(\mathbf{x}_i'\boldsymbol{\beta})\beta_j \quad (14.2)$$

其中，$F'(z)=\partial F(z)/\partial z$。正如任何非线性模型一样，边际效应会随计算点 $\mathbf{x}_i$ 不同而不同，同时因 $F(\cdot)$的不同选取而千差万别。表 14.3 的最后一列给出常用二值结果模型的边际效应。

非线性模型的边际效应已在 5.2.4 节讨论过。给定特定模型，存在几种计算平均边际效应的方法。一种最好的方法是，使用 $N^{-1}\sum_i F'(\mathbf{x}_i'\hat{\boldsymbol{\beta}})\hat{\beta}_j$，即边际效应的样本平均。不过，一些程序在回归元的样本均值处加以计算，即 $F'(\bar{\mathbf{x}}'\hat{\boldsymbol{\beta}})\hat{\beta}_j$。前面构造的测量是在 $\bar{y}$ 处，即 y 的样本均值处进行计算，所以 $F'(\mathbf{x}'\boldsymbol{\beta})=\bar{y}$ 且 $F'(\mathbf{x}'\boldsymbol{\beta})=F'(F^{-1}(\bar{y}))$。对 logit 模型来说，尤其简单，从而得到估计边际效应 $\bar{y}(1-\bar{y})\hat{\beta}_j$。对于特定模型的进一步讨论，将在 14.3.4 节至 14.3.7 节给出。

然而，许多研究只报告回归系数。标准的二值模型是单指标模型，因此，两个不同回归元的系数之比等于其边际效应之比。由于 $F'(\cdot)>0$，所以系数符号就给出边际效应的符号。系数能用于获得边际效应的上界。对于 logit 模型来说，$\partial p/\partial x_j\leqslant 0.25\hat{\beta}_j$，由于 $\Lambda(\mathbf{x}'\boldsymbol{\beta})(1-\Lambda(\mathbf{x}'\boldsymbol{\beta}))\leqslant 0.25$，所以当 $\Lambda(\mathbf{x}'\boldsymbol{\beta})=0.5$ 且 $\mathbf{x}'\boldsymbol{\beta}=0$ 时达到最大值。对于 probit 模型来说，$\partial p/\partial x_j\leqslant 0.4\hat{\beta}_j$，由于 $\phi(\mathbf{x}'\boldsymbol{\beta})\leqslant 1/\sqrt{2\pi}\simeq 0.4$，当 $\Phi(\mathbf{x}'\boldsymbol{\beta})=0.5$ 且 $\mathbf{x}'\boldsymbol{\beta}=0$ 时，达到最大值。

14.3.3 ML 估计

考察已知样本$(y_i,\mathbf{x}_i)$时的估计问题，$i=1,\cdots,N$，其中假定对于不同 i 具有独立性。结论是针对式(14.1)定义 p_i 给出的，对 logit 与 probit 设定的专门研究则稍后给出。

一般二值结果模型的 MLE

结果服从贝努利分布，而二项式分布仅仅是含有一种试验的情况。就 y_i 密度而言，一种非常方便的简洁记号，或更正式地讲，其概率质量函数(**probability mass function**)是：

$$f(y_i\mid \mathbf{x}_i)=p_i^{y_i}(1-p_i)^{1-y_i},\quad y_i=0,\ 1 \quad (14.3)$$

其中，$p_i=F(\mathbf{x}_i'\boldsymbol{\beta})$。从而，得到概率 p_i 与$(1-p_i)$，这是因为 $f(1)=p^1(1-p)^0=p$，而 $f(0)=p^0(1-p)^1=1-p$。

由密度(14.3)得出，对数密度 $\ln f(y_i)=y_i\ln p_i+(1-y_i)\ln(1-p_i)$。给定对应于不同 i 的独立性且关于 p_i 的模型式(14.1)，对数似然函数是：

$$\mathcal{L}_N(\boldsymbol{\beta})=\sum_{i=1}^{N}\{y_i\ln F(\mathbf{x}_i'\boldsymbol{\beta})+(1-y_i)\ln(1-F(\mathbf{x}_i'\boldsymbol{\beta}))\} \quad (14.4)$$

求关于 $\boldsymbol{\beta}$ 的导数，得出 MLE $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$是

$$\sum_{i=1}^{N}\left\{\frac{y_i}{F_i}F_i'\mathbf{x}_i-\frac{1-y_i}{1-F_i}F_i'\mathbf{x}_i\right\}=\mathbf{0}$$

的解，其中 $F_i=F(\mathbf{x}_i'\boldsymbol{\beta})$，$F_i'=F'(\mathbf{x}_i'\boldsymbol{\beta})$，而 $F'(z)=\partial F(z)/\partial z$。一旦对含有共同分

母 $F_i(1-F_i)$ 的分式进行变换，同时加以简化，得到 ML 一阶条件：

$$\sum_{i=1}^{N}\frac{y_i-F(\mathbf{x}_i'\boldsymbol{\beta})}{F(\mathbf{x}_i'\boldsymbol{\beta})(1-F(\mathbf{x}_i'\boldsymbol{\beta}))}F'(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i=\mathbf{0} \tag{14.5}$$

尽管 $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$ 没有显式解，但牛顿—拉夫森迭代法通常很快就收敛，因为至少对 probit 与 logit 模型来说，对数似然均是全局凹的。

MLE 的一致性

若已知 $\mathbf{x}$ 时 y 的条件密度被正确地设定，则 MLE 是一致的(**consistent**)。由于此处密度必是贝努利密度，所以唯一可能的错误设定是，贝努利概率被错误设定。因此，当 $p_i \equiv F(\mathbf{x}_i'\boldsymbol{\beta})$ 时，MLE 是一致的，否则是非一致的。

更正式地讲，注意到，二值数据 $\mathrm{E}[y]=1\times p+0\times(1-p)=p$。给定式(14.1)，得出：

$$\mathrm{E}[y_i|\mathbf{x}_i]=F(\mathbf{x}_i'\boldsymbol{\beta}) \tag{14.6}$$

它同样蕴含，一阶方程(14.5)的左边具有零期望值，即一致性的根本条件。倘若条件均值被正确设定，一致性的这个特殊结果对于 LEF 密度来说就成立(参见 5.7.3 节)，而贝努利密度是 LEF 密度。

MLE 的分布

已知正确设定密度，则有 $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}\overset{a}{\sim}\mathcal{N}[\boldsymbol{\beta},(-\mathrm{E}[\partial^2\mathcal{L}_N/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'])^{-1}]$(参见 5.6.4 节)。对于式(14.4)求关于 $\boldsymbol{\beta}'$ 的导数，并对期望值取负数，得到估计渐近方差矩阵(**asymptotic variance**)：

$$\hat{\mathrm{V}}[\hat{\boldsymbol{\beta}}_{\mathrm{ML}}]=\left(\sum_{i=1}^{N}\frac{1}{F(\mathbf{x}_i'\hat{\boldsymbol{\beta}})(1-F(\mathbf{x}_i'\hat{\boldsymbol{\beta}}))}F'(\mathbf{x}_i'\hat{\boldsymbol{\beta}})^2\mathbf{x}_i\mathbf{x}_i'\right)^{-1} \tag{14.7}$$

其中，由于 $\mathrm{E}[y_i-F(\mathbf{x}_i'\boldsymbol{\beta})]=0$，故可简化。这个方程矩阵具有简单形式 $(\sum_i\hat{w}_i\mathbf{x}_i\mathbf{x}_i')^{-1}$，这里的权数 $\hat{w}_i$ 已由式(14.7)给出。

由于一致性只要求对条件均值或概率正确设定，所以当然考察准 MLE(参见 5.7 节)，并将推断建立在方差矩阵的三明治形式 $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$ 基础上，而不是式(14.7)使用的 $-\mathbf{A}^{-1}$ 基础上。这里：

$$\mathrm{V}[y_i|\mathbf{x}_i]=F(\mathbf{x}_i'\boldsymbol{\beta})(1-F(\mathbf{x}_i'\boldsymbol{\beta})) \tag{14.8}$$

因为 $\mathrm{V}[y]=(1-p)^2\times p+(0-p)^2\times(1-p)=p(1-p)$。经过一些代数运算，可以证明，一旦假设对不同 i 具有独立性，这蕴含 $\mathbf{A}=-\mathbf{B}$，从而 $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}=\mathbf{A}^{-1}$。式(14.8)不成立的唯一方式是，当遭受更基本的非一致性问题时，出现 $p\neq F(\mathbf{x}'\boldsymbol{\beta})$ 时的情况。

二值结果模型很有特色，因为当数据对于不同 i 是独立的，利用三明治形式不存在什么优势。转向稳健方差矩阵估计的唯一原因是，因出现聚集，故观测值关于 i 是相关的，于是需要稳健估计，即甚至对于聚集是稳健的(参见 24.5 节)，而不是对条件方差的错误设定是稳健的。

14.3.4 logit 模型

logit 模型(**logit model**)或者逻辑斯蒂回归模型(**logistic regression model**)设定：

$$p=\Lambda(\mathbf{x}'\boldsymbol{\beta})=\frac{e^{\mathbf{x}'\boldsymbol{\beta}}}{1+e^{\mathbf{x}'\boldsymbol{\beta}}} \tag{14.9}$$

其中，$\Lambda(\cdot)$表示逻辑斯蒂 cdf(更详细内容参见 14.4.1 节)，而 $\Lambda(z)=e^z/(1+e^z)=1/(1+e^{-z})$。

由于 $\Lambda'(z)=\Lambda(z)[1-\Lambda(z)]$，故 logit 的 MLE 一阶条件(14.5)简化成：

$$\sum_{i=1}^{N}(y_i-\Lambda(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{x}_i=\mathbf{0} \tag{14.10}$$

因此，类似于 OLS 回归，原始残差 $y_i-\Lambda(\mathbf{x}_i'\boldsymbol{\beta})$与回归元是正交的。由于 $\Lambda(\cdot)$是贝努利密度的典型连结函数(**canonical link function**)(参见 5.7.4 节)，所以出现简单形式。

若回归元 $\mathbf{x}_i$ 包括截距，则式(14.10)蕴含$\sum_i(y_i-\Lambda(\mathbf{x}_i'\hat{\boldsymbol{\beta}}))=0$，因而 logit 残差和为 0。从而得出，样本内预测概率平均值 $N^{-1}\sum_i\Lambda(\mathbf{x}_i'\hat{\boldsymbol{\beta}})$一定等于样本频率 $\bar{y}$。

对于 logit 模型来说，其边际效应(**marginal effects**)相当容易地从系数中获得，因为$\partial p_i/\partial x_{ij}=p_i(1-p_i)\beta_j$，其中 ，$p_i=\Lambda_i=\Lambda(\mathbf{x}_i'\boldsymbol{\beta})$。一旦在 $p_i=\bar{y}$ 处进行计算，得到 $\bar{y}(1-\bar{y})\hat{\beta}_j$ 的边际效应大概估计。例如，对于 $0.3<p_i<0.7$，$\partial p_i/\partial x_{ij}$位于 $0.21\beta_j$ 与 $0.25\beta_j$ 之间。对于 $p_i\simeq 0.0$ 的数据来说，在此情况下，大多数结果为 0，$\partial p_i/\partial x_{ij}=p_i\beta_j$，因而 β_j 给出当 x_{ij}变化时关于 $y_i=1$ 概率的成比例效应。

在统计学文献中，对系数非常普遍的解释是依据关于优势比，而不是概率的边际效应来表述。对于 logit 模型来说，有：

$$\begin{aligned} p&=\exp(\mathbf{x}'\boldsymbol{\beta})/(1+\exp(\mathbf{x}'\boldsymbol{\beta}))\\ \Rightarrow\frac{p}{1-p}&=\exp(\mathbf{x}'\boldsymbol{\beta}) \\ \Rightarrow\ln\frac{p}{1-p}&=\mathbf{x}'\boldsymbol{\beta}\end{aligned} \tag{14.11}$$

这里，$p/(1-p)$测算 $y=1$ 的概率相对于 $y=0$ 的概率之比，并称为优势比[1](**odds ratio**)或相对风险(**relative risk**)。例如，考察药物研究，其中，$y=1$ 表示存活，而 $y=0$ 表示死亡，同时回归元包括用药量测量。优势比为 2 意味着，存活发生比是死亡的 2 倍。对于 logit 模型来说，对数优势比(**log-odds ratio**)关于回归元是线性的。

一些条件分析及软件包都运用式(14.11)的第二个等式。假定第 j 个回归元增加一个单位。那么，$\exp(\mathbf{x}'\boldsymbol{\beta})$增大到 $\exp(\mathbf{x}'\boldsymbol{\beta}+\beta_j)=\exp(\mathbf{x}'\boldsymbol{\beta})\times\exp(\beta_j)$。由式(14.11)可得，优势比增加 $\exp(\beta_j)$倍。因而，例如，对于 logit 模型来说，0.1 的斜率

〔1〕 又称为发生比。——译者注

参数意味着，回归元上增加一个单位，最初优势比会增加 $\exp(0.1)\simeq 1.105$ 倍。这是用增大 0.105 比例乘以最初优势比，因此，生存的相对概率提高 10.5%。logit 模型的这一解释广泛用于生物统计学应用之中。

对于经济学家来说，一种更自然的方式是，将式(14.11)的第二个等式或第三个等式解释成 β_j 是半弹性(**semi-elasticity**)的含义。然后，采用微分方法，将 logit 模型斜率 0.1 参数解释成回归元上增加一个单位会使优势比增大 0.1 倍。对于非常小的 β_j 来说，这与统计学中所使用的解释完全一样，从而 $\exp(\beta_j)-1\simeq\beta_j$。

14.3.5 *probit 模型*

probit 模型(**probit model**)将条件概率设定成：

$$p=\Phi(\mathbf{x}'\boldsymbol{\beta})=\int_{-\infty}^{\mathbf{x}'\boldsymbol{\beta}}\phi(z)dz \tag{14.12}$$

其中，$\Phi(\cdot)$表示标准正态 cdf，其导数 $\phi(z)=(1/\sqrt{2\pi})\exp(-z^2/2)$，它是标准正态密度函数。

probit MLE 的一阶条件是：

$$\sum_{i=1}^{N}w_i(y_i-\Phi(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{x}_i=\mathbf{0}$$

这与 logit 模型不同，其中，权数 $w_i=\phi(\mathbf{x}_i'\boldsymbol{\beta})/[\Phi(\mathbf{x}_i'\boldsymbol{\beta})(1-\Phi(\mathbf{x}_i'\boldsymbol{\beta}))]$ 随观测值而变化。

probit 模型的边际效应是，$\partial p_i/\partial x_{ij}=\phi(\mathbf{x}_i'\boldsymbol{\beta})\beta_j=\phi(\Phi^{-1}(p_i))\beta_j$，其中 $p_i=\Phi(\mathbf{x}_i'\boldsymbol{\beta})$。尽管$\partial p_i/\partial x_{ij}\leqslant 0.40\beta_j$，但由于 $\phi(z)\leqslant\phi(0.5)=1/\sqrt{2\pi}$，故不存在类似于 logit 模型的进一步简化。

probit 模型不像 logit 模型那样简单。不过，若起点是潜正态回归模型(参见 14.4 节)，则因为它是一个自然的模型仍会被广泛运用。

14.3.6 *OLS 估计*

一种对 logit 或 probit 的可供选择是，y 对 $\mathbf{x}$ 的 OLS 回归(**OLS regression**)。这具有明显的缺陷，可能出现所求的预测概率 $\mathbf{x}_i'\hat{\boldsymbol{\beta}}$ 是负的或大于 1。

不过，OLS 估计量作为解释工具仍是有用的。在实际应用中，它提供了当 x 变化时样本平均对 $y=1$ 的概率的边际效应的一种合理的直接估计值，尽管它关于个体概率提供了不好的模型。实际上，它提供哪些变量是统计显著的一种良好指南。在许多应用中，可以证明，对于所有样本观测值来说，$0<\mathbf{x}_i'\hat{\boldsymbol{\beta}}<1$，在此情况下，OLS 就更为合理。

倘若运用 OLS 估计量，则因异方差性(**heteroskedasticity**)而应对标准误差加以校正。当概率 $p_i=\mathbf{x}_i'\boldsymbol{\beta}$时，可判定线性回归是正确的。于是，$y_i|\mathbf{x}_i$ 具有均值$\mathbf{x}_i'\boldsymbol{\beta}$，而异方差方差 $\mathbf{x}_i'\boldsymbol{\beta}(1-\mathbf{x}_i'\boldsymbol{\beta})$会随 $\mathbf{x}_i$ 而变化。

原则上，当 $p_i=\mathbf{x}_i'\boldsymbol{\beta}$ 时，可能获得一种更有效的 ML 估计。由式(14.5)知，ML 一阶条件是 $\sum_i\mathbf{x}_i(y_i-\mathbf{x}_i'\boldsymbol{\beta})/[\mathbf{x}_i'\boldsymbol{\beta}(1-\mathbf{x}_i'\boldsymbol{\beta})]=\mathbf{0}$。由于此估计量将非常大的权数

施加给接近于 0 或 1 的含有 $\mathbf{x}_i'\boldsymbol{\beta}$ 的观测值，故它在数值形式上表现得不稳定。另外，与 OLS 相比，有效性提高往往不大。

即使含有异方差标准误差的 OLS 估计可作为一种解释数据分析的有益工具，但最好是对最终数据分析使用 logit 或 probit MLE。

14.3.7 *选择二值模型*

应该运用哪一个模型呢？是 logit 还是 probit？该问题是本节要探讨的。

理论上的考虑

从理论上讲，回答要依赖于数据生成过程，而数据生成过程却是未知的。与 ML 的其他一些应用不同，在设定分布上不存在什么问题，即关于(0，1)变量的唯一可能分布是贝努利分布。该问题依赖于对此分布参数的函数形式的设定。若数据生成过程具有 $p=\Lambda(\mathbf{x}'\boldsymbol{\beta})$，则应该使用 logit 模型，而建立在其他模型诸如probit 基础上的一些估计量均潜在地是非一致的。不过，若数据生成过程具有$p=\Phi(\mathbf{x}'\boldsymbol{\beta})$，则类似的定性结论仍然成立，在此情况下，应该使用 probit 模型。$p=\mathbf{x}'\boldsymbol{\beta}$ 是最不可能出现的，因为那样 p 没有限制在 0 与 1 之间。

可是，模型错误设定的理论后果并没有如此之大。如果回归元具有分布，使得以线性组合 $\mathbf{x}'\boldsymbol{\beta}$ 为条件的每个回归元的均值关于 $\mathbf{x}'\boldsymbol{\beta}$ 是线性的，那么可以证明，选择错误函数 F 同样会影响到所有斜率参数，以致斜率参数之比对于不同模型都是常值；参见鲁德(1983)。借助于球面分布族，包括多变量正态分布，该条件会得到满足。

就一阶条件与渐近分布而言，logit 模型拥有相对简单的形式。伯克森(Berkson，1951)曾经推广 logit 模型，给出与最初 probit 模型相比他更偏爱 logit 模型的几个原因之一。在广泛用于生物统计学的广义线性模型框架下，logit 模型是一个自然的模型，因为它对应于使用二项分布的最简洁的联系。依据对数优势比对系数的解释也是 logit 模型的吸引人之处。

迄今为止，使用 logit 模型的另一个动机是判别分析(**discriminant analysis**)。在判别分析中，y 是随机变量，x 也是随机变量，x 是可观测的，y 却不是可观测的。给定 x 时，我们需要决定 y 是否等于 0 或 1。一个经典例子是，对人类($y=0$ 或 1)头盖骨进行分类，即什么类型属于给定头盖骨的各种维度。当已知 y 时特征 x 的条件分布服从多变量正态分布时，已知 x 时 y 的后验概率就类似于 logit 模型的概率。对于更详细内容，参见雨宫(Amemiya，1981，第 1 507～1 510 页)以及马达拉(Maddala，1983，第 17～21 页)。

与之相比，probit 模型因潜正态随机变量激励而拥有引人注目之处(参见 14.4 节)，同时很自然地推广到 Tobit 模型(参见第 16 章)。正是由于这些原因，许多经济学家都运用 probit 模型。

经验上的考虑

从经验上讲，不是运用 logit 就是运用 probit。出自 probit 模型的预测概率与出自 logit 模型的预测概率之间，通常存在极小差异。在概率接近于 0 或 1 的尾部时，它们之间的差异最大。若关注内容只是在于样本，而不是每一个个体的边际效应，则其差别就相当小。

用于比较模型的一个距离是拟合对数似然，这是因为给定关于 p_i 模型，存在着一致认识：对数似然是正确的观点，而且 logit 模型与 probit 模型拥有相同个数的参数。因而，对于每一个模型，计算：

$$\mathcal{L}_N(\hat{\boldsymbol{\beta}}) = \sum_i \{y_i \ln \hat{p}_i + (1-y_i)\ln(1-\hat{p}_i)\}$$

其中，$\hat{p}_i = \Lambda(\mathbf{x}_i'\hat{\boldsymbol{\beta}}_{\text{logit}})$或 $\hat{p}_i = \Phi(\mathbf{x}_i'\hat{\boldsymbol{\beta}}_{\text{probit}})$。这两个模型的拟合对数似然往往是非常相似的，再次表明，使用一个模型而不是另一个模型带来的额外好处很少。关于更正式非嵌套模型检验内容，可参见佩萨兰和佩萨兰（Pesaran and Pesaran，1995）以及 8.5 节。

各种不同的模型会产生截然不同的回归参数的 $\hat{\beta}$ 估计值。然而，这只是利用各种不同概率公式的人为现象。更有意义的是，去比较不同模型的边际效应，因为这种测量对于三种模型来说具有类似标度。由 14.2.3 节知，对于 logit 模型，$\partial p/\partial x_j \leqslant 0.25\hat{\beta}_j$。对于 probit 模型，$\partial p/\partial x_j \leqslant 0.4\hat{\beta}_j$，而对于 OLS 来说，$\partial p/\partial x_j = \hat{\beta}_j$。从而，提出一个经验法则（**rule of thumb**）：

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{logit}} &\simeq 4\hat{\boldsymbol{\beta}}_{\text{OLS}} \\ \hat{\boldsymbol{\beta}}_{\text{probit}} &\simeq 2.5\hat{\boldsymbol{\beta}}_{\text{OLS}} \\ \hat{\boldsymbol{\beta}}_{\text{logit}} &\simeq 1.6\hat{\boldsymbol{\beta}}_{\text{probit}}\end{aligned} \tag{14.13}$$

雨宫（Amemiya，1991，第 1 488 页）已经证明，当 $0.1 \leqslant p \leqslant 0.9$ 时，这些比较关系对斜率参数相当奏效。较大偏离出现在各个不同模型的尾部。对于 logit 模型来说，稍后将给出建立在式（14.18）基础上的一种可供选择的方法；使用 $\hat{\boldsymbol{\beta}}_{\text{logit}} \simeq (\pi/\sqrt{3})\hat{\boldsymbol{\beta}}_{\text{probit}}$。

内生回归元

可对 logit 与 probit 模型加以推广，用以处理微观经济计量分析中普遍出现的许多复杂情况。特别地，内生回归元可利用类似于 16.8.2 节给出的关于删失数据的那些方法以及将在第 23 章阐述的面板数据方法都可用于分析内生回归元。

对于这类复杂情况，以线性概率模型加以研究比较容易，倘若标准误差对异方差性可调整，从而应用标准线性模型方法。即使最终运用 logit 与 probit 模型，对于解释性分析来说，线性模型也是有益的。

14.3.8 *确定模型适合性*

关于非线性模型的模型诊断与选择，已在 8.7 节阐述。这里，考察对二值结果模型的专门研究。不存在单个最佳测量，因此，统计软件包会报告雨宫（Amemiya，1981）与马达拉（Maddala，1983）曾详述过的几种测量。

伪 R^2

在线性回归模型中，标准拟合优度是 R^2。而对非线性模型的推广称为伪 R^2（**pseudo-R^2**），它有几种可能的推广形式。

更受喜欢的测量是 8.7.1 节记为 R^2_{RG}的相对增益测量。这种测量并不总是可以计算的，但它适合于二值结果模型，因为 $Q_{\max}$即对数似然的最大可能值为零。为

了获得此结果，注意到，最佳可能拟合显然是 y^*，它以概率 $p=1$ 预测 $y=1$ 而以概率 $1-p=0$ 预测 $y=0$，在这种情况下，$f(y^*)=1$ 且 $\ln f(y^*)=0$。于是，$R^2_{RG}=1-(0-Q_{fit})/(0-Q_0)=1-Q_{fit}/Q_0$。从而，得出由麦克法登（McFadden，1974）提出的关于二值结果模型的 R^2 测量：

$$R^2_{二值}=1-\frac{\mathcal{L}_N(\hat{\boldsymbol{\beta}})}{\mathcal{L}_N(\bar{y})} \tag{14.14}$$

$$=1-\frac{\sum_i[y_i \ln \hat{p}_i+(1-y_i)\ln(1-\hat{p}_i)]}{N[\bar{y}\ln\bar{y}+(1-\bar{y})\ln(1-\bar{y})]}$$

其中，$\hat{p}_i=F(\mathbf{x}_i'\hat{\boldsymbol{\beta}})$，而 $\bar{y}=N^{-1}\sum_i y_i$。

针对许多特定的二值数据，另一些关于 R^2 的测量已由雨宫（Amemiya，1981）与马达拉（Maddala，1983）给出。一种明显的测量结果是，y_i 与 $\hat{p}_i$ 之间样本相关系数的平方。这些额外测量之一，也要归功于麦克法登，而且许多参考文献都给出这个测量值而不是式（14.14）的 R^2。

预测结果

在线性回归模型中，拟合优度经常通过拟合值与实际值的比较来计算。对于二值数据来说，拟合值 $\hat{y}$ 应是二值的，因为 y 是二值的。准则 $\sum_i(y_i-\hat{y}_i)^2$ 会给出错误预测的数，若$(y, \hat{y})$等于$(1, 0)$或$(0, 1)$，则会出现此情况。一个明显的预测规则是，当 $\hat{p}=F(\mathbf{x}'\hat{\boldsymbol{\beta}})>0.5$ 时，设 $\hat{y}=1$。不过，这有一个弱点，即当样本大部分满足 $y=1$ 时，常常有 $\sum_i(y_i-\hat{y}_i)^2=n(1-\bar{y})$，因为很可能 $\hat{p}>0.5$，因此，对于所有观测值 $\hat{y}=1$。当样本大部分满足 $y=0$ 时，会出现类似问题。

更一般地，考察截止值范围。当 $\hat{p}>c$ 时，设定 $\hat{y}=1$，我们得到受试者工作特性（**receiver operating characteristics**，ROC）曲线[1]，它画出当截断值（**cutoff value**）c 改变时，$y=1$ 值正确分类部分与 $y=0$ 值错误分类部分对于 $c=1$ 来说，所有值均预测成为 1，因而所有 $y=1$ 值是正确分类的，而所有 $y=0$ 值却错误分类，从而 ROC 曲线取值$(0, 0)$。类似地，对于 $c=0$ 来说，ROC 曲线取值$(1, 1)$。

---

〔1〕受试者工作特性曲线，又称为接收者操作特性曲线。ROC 分析起源于 20 世纪 50 年代的统计决策理论。后来，应用于雷达信号观察能力的评价，20 世纪 60 年代中期，有大量成功用于实验心理学和心理物理学研究。勒斯蒂德（Lusted）首次提出了 ROC 分析可用于医学决策评价。自从 20 世纪 80 年代起，该方法广泛用于医学诊断性能的评价。

ROC 曲线用于二分类判别效果的分析与评价，一般自变量为连续变量，因变量为二分类变量。基本原理是：通过截止点（cutoff point/cutoff value，分界值或决定阈）的移动，获得多对灵敏度（sensitivity）和误判率[1-Specificity（特异度）]，以灵敏度（真阳性率）为纵轴标，以误判率（假阳性率）为横轴标，连接各点绘制曲线，然后计算曲线下的面积，面积越大，判断价值越高。其中，灵敏度表示把实际真值判断为真值的概率；特异度表示把实际的假值判断为假值的概率；误判率表示把实际的假值判断为真值的概率，其值等于 1－特异度。

将绘成的曲线与 45°直线对比，若差不多重合，说明自变量对因变量的判断价值很差，若越远离 45°直线，即曲线下的面积越大，说明自变量对因变量的判断价值越好，即根据自变量可以较为正确地判断因变量。

目前，ROC 曲线在医学诊断中广泛运用。传统的诊断试验评价方法有一个共同特点，必须将试验结果分为两类，再进行统计分析。ROC 曲线的评价方法与传统的评价方法不同，无须此限制，而是根据实际情况，允许有中间状态，可将试验结果划分为多个有序分类，如正常、大致正常、可疑、大致异常和异常五个等级再进行统计分析。——译者注

若模型没有预测能力，则 ROC 曲线是这些点之间的直线。该曲线越弯曲，同时它下面区域越大，则模型预测力就越好。

预测概率

由于二值数据服从简单的离开分布，一种明显的方法是，将 $y=1$ 的样本平均预测概率与样本频率 $N^{-1}\sum_i \hat{p}_i$ 加以比较，其中，$\hat{p}_i=F(\mathbf{x}_i'\hat{\boldsymbol{\beta}})$，样本频率为 $\bar{y}$。不过，对于具有截距模型来说，这没有什么用途，因为当 ML 一阶条件蕴含 $\sum_i[y_i-\Lambda(\mathbf{x}_i'\hat{\boldsymbol{\beta}})]=0$ 时，$N^{-1}\sum_i \hat{p}_i=\bar{y}$ 总是成立的。对于通过 OLS 进行估计的情况，类似结论成立，就 probit 模型而言，此结论并不准确，但实际上却相当接近。

然而，这一方法能用于对子样本的预测，然后建立 8.2.6 节给出的卡方拟合优度检验的基础。

14.4 潜变量模型

潜变量[1]（**latent variable**）是指不完全观测到的变量。潜变量会以两种不同方式引入二值结果模型中。第一种方式中的潜变量是指，关注事件发生的不可观测到的倾向。第二种方式中的潜变量是指，关注事件发生出现时效用上的差异（差），这里假定二值结果是个体选择的结果。显然，后一种方法需要在下述两种回归元之间加以区分，即对于给定个体来说，随不同可供选择而变化的回归元与给定个体随不同可供选择而不变的回归元，诸如社会经济特征。

应该强调的是，如同 14.3 节一样，二值结果服从贝努利分布。潜变量模型只对贝叶斯参数的特定函数形式提供了一个理论框架。

潜变量模型可被推广到多项式结果与删除结果（第 15 章和第 16 章将阐述）。潜变量模型同样提供利用增广数据进行贝叶斯分析（参见 13.7 节）。二值数据与多项式数据的贝叶斯分析的简要讨论将在 15.7.2 节和 15.8.2 节给出。

14.4.1 *指标函数模型*

在指标函数（**index function**）公式中，关注内容在于解释基本不可观测连续随机变量，但我们观测到的全部内容是二值变量 y，y 依据 y^* 是否经过门限值而取值为 1 或 0。y^* 的不同分布会导致各种不同的二值结果模型。

设 y^* 表示潜变量（或不可观测变量），诸如对劳动力供给进行建模时的工作意愿去向。y^* 的一个回归模型是指标函数模型（**index function model**）：

$$y^*=\mathbf{x}'\boldsymbol{\beta}+u \tag{14.15}$$

不过，当 y^* 不可观测时，就不能估计这个模型。相反，我们观测到：

$$y=\begin{cases}1, & \text{当 } y^*>0\\ 0, & \text{当 } y^*\leqslant 0\end{cases} \tag{14.16}$$

其中，门限值 0 是下述将要解释的正规化。

---

[1] 又称为隐变量。——译者注

已知式(14.16),有:

$$\begin{aligned}\Pr[y=1|\mathbf{x}]&=\Pr[y^*>0]\\&=\Pr[\mathbf{x}'\boldsymbol{\beta}+u>0]\\&=\Pr[-u<\mathbf{x}'\boldsymbol{\beta}]\\&=F(\mathbf{x}'\boldsymbol{\beta})\end{aligned}\tag{14.17}$$

其中,F 表示 $-u$ 的 cdf,在密度关于 0 对称的通常情况下,F 等于 u 的 cdf。

因此,指标函数模型给出式(14.1)中 $F(\cdot)$ 函数形式的动机。

probit 模型和 logit 模型

若误差 u 服从标准正态分布,则是 probit 模型,从而由式(14.17)得到,$\Pr[-u<\mathbf{x}'\boldsymbol{\beta}]=\Phi(\mathbf{x}'\boldsymbol{\beta})$,其中,$\Phi(\cdot)$表示标准正态的 cdf。

现在,引进逻辑斯蒂分布(**logistic distribution**)。在其标准形式中,逻辑斯蒂分布的 cdf 为:

$$\Lambda(u)=e^u/(1+e^u),\quad -\infty<u<\infty\tag{14.18}$$

其密度函数 $\Lambda'(u)=e^u/(1+e^u)^2$ 关于 0 是对称的,并且逻辑斯蒂随机变量均值为 0,且方差为 $\pi^3/3\simeq1.184^2$。

当误差 u 服从逻辑斯蒂分布,即 logit 模型,由式(14.17)得到,$\Pr[-u<\mathbf{x}'\boldsymbol{\beta}]=\Lambda(\mathbf{x}'\boldsymbol{\beta})$。注意到,这两个模型中因为 $\mathrm{V}[u]$不同,所以 $\boldsymbol{\beta}$表示不同的标度。

识别考虑

单指标模型的识别(**identification**)要求对 u 的方差进行限制,因为单指标模型仅能识别 $\boldsymbol{\beta}$,至多差一个常值标度。所能观测到的全部内容是,y^* 是否>0 或等价地是否 $\mathbf{x}'\boldsymbol{\beta}+u>0$。可是,这等价于 $\mathbf{x}'\boldsymbol{\beta}^++u^+>0$,其中,$\boldsymbol{\beta}^+=a\boldsymbol{\beta}$以及 $u^+=au$,对于任何 $a>0$。如果对误差的方差(u 或 u^+)施加约束,便确保 $\boldsymbol{\beta}$的唯一性。在 probit 模型中,设该误差方差为 $\pi^2/3$。

指标模型的门限不必是 0。更一般地,如果当 $y^*>\mathbf{z}'\boldsymbol{\delta}$ 时 $y=1$,那么式(14.17)变成 $\Pr[y=1]=F(\mathbf{x}'\boldsymbol{\beta}-\mathbf{z}'\boldsymbol{\delta})$,于是,$\boldsymbol{\delta}$ 能单独识别,当且仅当 $\mathbf{z}$ 的所有分量与 $\mathbf{x}$ 的所有分量均不一样。特别地,若 $\mathbf{x}$ 包含截距,$\mathbf{z}$ 也包含截距,则这些不能单独进行识别,所以要对门限截距正规化为 0。注意到,误差分布均值也需要加以正规化。就 logit 模型与 probit 模型而言,把它设为 0。

讨论

指标函数模型蕴含对 $\boldsymbol{\beta}$的直接解释,即当 $\mathbf{x}$ 变动一个单位时潜变量 y^* 上的变化。即使 y^* 是不可观测的,但人们使用 u 的设定方差的知识,这种解释是有意义的。例如,在 probit 模型中,斜率参数 0.5 意味着回归元上的单位变动会导致 y^* 上的 0.5 个标准差变化,因为在这一模型中,y^* 的方差等于 1。

指标函数方法广泛运用的一种推广(**extension**)是,有序离散选择模型(参见 15.9 节),以及关于删失样本及选择样本的模型(参见第 16 章)。

14.4.2 *随机效用模型*

在随机效用公式中,消费者在 0 与 1 之间进行选择,依据是哪一种选取具有较

高的满意度或效用。若选项 1 具有较高效用，则离散变量 y 取值 1，而若选项 0 具有较高效用，则 y 取值 0。

可加随机效用模型(ARUM)对选项 0 与 1 设定成：

$$U_0 = V_0 + \varepsilon_0 \qquad (14.19)$$
$$U_1 = V_1 + \varepsilon_0$$

其中，V_0 与 V_1 均表示效用的确定性成分，而 ε_0 与 ε_1 均表示效用的随机性成分。一个简单的例子是 $V_0 = \mathbf{x}'\boldsymbol{\beta}_0$ 且 $V_1 = \mathbf{x}'\boldsymbol{\beta}_1$，具有较高效用的选项被选取。

具有较高效应的选项被选取。比如说，当 $U_1 > U_0$ 时，我们观测到 $y=1$。由于效用的随机性成分存在，所以这是一个满足

$$\begin{aligned}\Pr[y=1] &= \Pr[U_1 > U_0] \\ &= \Pr[V_1 + \varepsilon_1 > V_0 + \varepsilon_0] \\ &= \Pr[\varepsilon_0 - \varepsilon_1 < V_1 - V_0] \\ &= F(V_1 - V_0)\end{aligned} \qquad (14.20)$$

的随机事件，其中，F 表示$(\varepsilon_0 - \varepsilon_1)$的 cdf。当 $V_1 - V_0 = \mathbf{x}'\boldsymbol{\beta}$ 时，得到 $\Pr[y=1] = F(\mathbf{x}'\boldsymbol{\beta})$。

由于当 $U_1 > U_0$ 时，有 $aU_1 > aU_0$，所以 ARUM 需要对标度进行正规化。这通常是借助于设定 $\varepsilon_0 - \varepsilon_1$ 的方差或 ε_0 与 ε_1 的方差来完成。

对 ε_0 与 ε_1 的分布进行各种不同设定，会给出不同的 $F(\cdot)$，从而得到各种离散选择模型。随机效用公式尤其对设定无序多项式选择模型有用(参见 15.5 节)。

probit 模型与 logit 模型

对式(14.19)误差分布的一种明显选择是，ε_0 与 ε_1 均服从正态分布。于是，$(\varepsilon_0 - \varepsilon_1)$服从正态分布。若对$(\varepsilon_0 - \varepsilon_1)$的方差进行正规化为 1，则得到 probit 模型，从而式(14.20)的 $F(\cdot)$是标准正态 cdf。

现在引入第 1 类极值分布(**type 1 extreme value distribution**)或对数威布尔分布(**log Weibull distribution**)。于是，随机变量 ε 具有密度：

$$f(\varepsilon) = e^{-\varepsilon}\exp(-e^{-\varepsilon}), \quad -\infty < \varepsilon < \infty \qquad (14.21)$$

而且 cdf $F(\varepsilon) = \exp(-e^{-\varepsilon})$。极值分布极少在经济计量学中应用，它可作为从相同分布抽取的 N 个随机变量的最大值在 $N \to \infty$ 时的极限分布。第 1 类极限分布是如下的特殊情况：在$(-\infty, \infty)$上拥有-2 与 5 之间的大部分质量是右偏斜的。它具有中位数$-\ln(-\ln(0.5)) \simeq 0.366\,51$，均值 $\Gamma'(1) \simeq 0.577\,22$，其中，$\Gamma(x)$表示伽玛函数的导数，而且方差 $\pi^2/6 \simeq 1.282\,55^2$。此分布可由对数正态来很好地逼近。

若假定 ε_0 与 ε_1 服从独立的第 1 类极限分布，就是 logit 模型。可以证明，其差服从逻辑斯蒂分布[参见约翰逊和科茨(Johnson and Kozt, 1970)]，所以式(14.20)中的 $F(\cdot)$是逻辑斯蒂 cdf。

作为这个结果的一种可供选择的推导是极值分布直接进行，稍后在 14.8 节给出。当 ARUM 被扩展到 15.5 节中在三个或更多可供选项之中选择的情况时，推导求出概率闭形式解是极其困难的。甚至在不存在闭形式解时，最近的计算发展

使得估计变得容易。

14.4.3 随可供选项变化的回归元

在绝大多数二值选择模型的应用中，有些回归元会随不同个体而变化，但一些回归元不一定会随可供选项而变化。

在一种极端情形下，回归元并不随可供选项而变化。例如，在决策参加工作的劳动力供给模型中，社会经济特征诸如收入与性别并不随可供选项而变化。一种潜在的回归元譬如工资率没有随着工作或不工作的选项而变化，但通常不包括这种回归元，因为它仅对那些选择工作的人来说是可观测的。

另外一种极端情形下，所有回归元可以随可供选项而变化。例如，在运输方式选择模型中，回归元可能是时间成本与两种运输模型的货币成本。

一般的混合 ARUM 是将式(14.19)中效用的确定性成分定义成：

$$V_{ij}=\mathbf{z}_{ij}'\boldsymbol{\alpha}_j+\mathbf{w}_i'\boldsymbol{\gamma}_j,\quad j=0,1 \tag{14.22}$$

其中，$\mathbf{z}_{ij}$ 表示随两个可供选项而取不同值的回归元，而 $\mathbf{w}_i$ 表示并不随选取而变化的个体特征。于是，由式(14.20)，得到：

$$\Pr[y_i=1]=F(\mathbf{z}_{i1}'\boldsymbol{\alpha}_1-\mathbf{z}_{i0}'\boldsymbol{\alpha}_0+\mathbf{w}_i'(\boldsymbol{\gamma}_1-\boldsymbol{\gamma}_0))$$

对于随可供选项不变的回归元(**alternative-invariant regressors**)来说，唯一的参数差($\boldsymbol{\gamma}_1-\boldsymbol{\gamma}_0$)是可以识别的。对于随可供选项且随个体而变化的回归元(**alternative-varying regressors**)来说，其系数会随可供选项不同而变化，但一种习惯做法是，令 $\boldsymbol{\alpha}_1=\boldsymbol{\alpha}_0=\boldsymbol{\alpha}$。例如，由旅行成本增加 1 美元引起的效用损失被认为是随各种不同运输方式而一样的，因而，ARUM 会导致：

$$\Pr[y_i=1]=F((\mathbf{z}_{i1}-\mathbf{z}_{i0})'\boldsymbol{\alpha}+\mathbf{w}_i'(\boldsymbol{\gamma}_1-\boldsymbol{\gamma}_0)) \tag{14.23}$$

这是最初二值选择模型(14.1)，其中一些回归元是随可供选项不变的回归元 $\mathbf{w}$，以及可供选择的不同回归元 $\mathbf{z}$ 的项之差。

14.5 基于选择的样本

每当样本选取部分地通过因变量 y 取值，而不是完全随机或部分基于由 $\mathbf{x}$ 的取值而决定的时候，就是基于选择抽样(**choice-based sampling**)。

一些离散数据模型均是重要的例子，因为调查经常故意对很少发生的选择进行过度抽样。例如，如果很少人选择通过公交车经常往来两地，就可能对乘公交车的人进行过度抽样。在医学文献中，对于病例对照分析[1](**case-control analysis**)来说，会出现同样问题，例如，二值数据分析可建立在那些患有心脏病发作的完全样本与具有相似特征且没有患心脏病发作的人的子样本基础上。基于选择抽样标准术语很少会使人误入歧途，因为它不是由个体选择而产生的。

[1] 又称为个案控制研究。——译者注

为了理解标准二值选择方法的非一致性，考察 logit 模型中唯一回归元是截距时的估计。于是，$\Lambda(\mathbf{x}_i'\boldsymbol{\beta})=\Lambda(\boldsymbol{\beta})$，同时 logit 的 MLE 一阶条件变成 $N^{-1}\sum_i\times(y_i-\Lambda(\boldsymbol{\beta}))=0$，所以 $\hat{\beta}=\ln(\bar{y}/(1-\bar{y}))$ 。很明显，$\hat{\beta}$ 的一致性需要随机样本，例如，因为对 $y=1$ 进行过度抽样会导致对 $\bar{y}$ 的过度估计，从而引起对 $\hat{\beta}$ 的过度估计。

获得给定内生抽样，诸如基于选择抽样时一致估计的方法，将在 24.4 节详细阐述。当已知过度抽样程度时，分析就简单易行。设 Q_1 表示总体中满足 $y=1$ 的部分，而 $H_1=\bar{y}$ 表示样本中满足 $y=1$ 的部分。类似地，定义 $Q_0=1-Q_1$ 且 $H_0=1-H_1$。于是，利用由曼斯基和莱尔曼（Manski and Lerman，1977）提出的一种加权 MLE(**weighred MLE**)可进行一致估计。对于二值结果模型来说，这对加权对数似然

$$\mathcal{L}_N^W(\boldsymbol{\beta})=\sum_{i=1}^{N}\left\{\left(\frac{Q_1}{H_1}\right)y_i\ln F(\mathbf{x}_i'\boldsymbol{\beta})+\left(\frac{Q_0}{H_0}\right)(1-y_i)\ln(1-F(\mathbf{x}_i'\boldsymbol{\beta}))\right\}$$

求极大值。例如，当结果 $y=1$ 被过度抽样时，$Q_1/H_1<1$，从而满足 $y=1$ 的过度抽样观测值均将被降低权数。这种估计量很容易利用允许对观测值加权的二值结果模型的任何程序来执行。于是，满足 $y=1$ 的观测值被赋予权数 Q_1/H_1，而满足 $y=0$ 的观测值则被赋予权数 Q_0/H_0。

雨宫（Amemiya，1985，9.5 节）给出关于二值与多项式数据的基于选择抽样的 ML 方法的详细归纳总结，包括当 Q_1 与 Q_0 是未知的时候。尽管加权 MLE 是无效的，但它实施简单且有效性损失可能不大。曼斯基和麦克法登（Manski and McFadden，1981a）已经提出一种更为有效的变形方法[参见雨宫和翁（Amemiya and Vuong，1987）]。科塞尔特（Cosselett，1981a，b）曾经提出完全有效的进一步精炼，但实施起来意义不大。英伯斯（Imbens，1992）以及兰开斯特和英伯斯（Lancaster and Imbens，1996）均提出了 GMM 估计作为一种可供选择的方法，它实施起来简单易行且完全有效。京和曾（King and Zeng，2001）给出二值 logit 模型的归纳总结；此外，他们考察当关注总体概率以低概率发生时小样本修正所引起的差异，甚至过度抽样。进一步详细内容，参见 24.4 节。

流行病学文献关注病例对照研究的 logit 模型。该方法归功于普伦蒂斯和派克（Prentice and Pyke，1979）。参见布雷斯洛（Breslow，1996），尤其是他的 4.3 节曾讨论经济计量学和流行病学文献之间的关系。

14.6 分组数据与加总数据

在一些应用中，只有分组数据或加总数据可以利用，但人们认为，个体特征可通过二值选择模型对其进行建模。当分组是建立在回归元的唯一值基础之上时，进行分组并不会引起什么问题，而且对回归元的每一个值而言存在许多观测值。在转向更现实问题之前，我们以这种简单的例子开始。

14.6.1 *伯克森最小卡方估计量*

假定回归元向量 $\mathbf{x}_i$ 只取 T 个不同的值，$i=1,\cdots,N$，其中，与 N 相比，T 更小

一些。于是，对回归元每个值来说，我们拥有关于 y 的多重观测值。这类分组数据称为每单元多观测值(**many observations per cell**)。特别地，在 $\mathbf{x}$ 具有低维数的实验数据中能出现此情况，并且是通过实验设计成为仅仅很少几个值的集合。设 $\mathbf{x}_t$ 表示 T 个不同的值，而 N_t 表示 x 第 t 个值的关于 y_t 的观测值个数，$t=1,\cdots,T$，因此有 $\sum_{t=1}^{T}N_t=N$，$\bar{p}_t$ 表示当 $\mathbf{x}_i=\mathbf{x}_t$ 时 $y_i=1$ 出现的次数。注意，下标 t 用于表示分组而并不一定表示时间。

对于满足 $\mathbf{x}_i=\mathbf{x}_t$ 的个体来说，如前所述，贝努利概率是：

$$p_t=\Pr[y_i=1\,|\,\mathbf{x}_i=\mathbf{x}_t]=F(\mathbf{x}_t'\boldsymbol{\beta}) \tag{14.24}$$

当对式(14.24)求反函数时，得到：

$$F^{-1}(p_t)=\mathbf{x}_t'\boldsymbol{\beta}$$

现在，p_t 未知，却能通过 $\bar{p}_t$ 加以估计，所以伯克森(Berkson)提出将 $F^{-1}(\bar{p}_t)$ 对 $\mathbf{x}_t$ 进行回归。因而，通过 LS 变换模型：

$$F^{-1}(\bar{p}_t)=\mathbf{x}_t'\boldsymbol{\beta}+v_t,\quad t=1,\cdots,T \tag{14.25}$$

加以估计。误差项 $v_t=F^{-1}(\bar{p}_t)-F^{-1}(p_t)$ 是异方差的，当 N_t 增大时其方差减小，从而 $\bar{p}_t$ 是 p_t 的一个较好估计值，同时还将依赖于 $F(\cdot)$ 的形状。由泰勒级数展开式[参见雨宫(Amemiya，1981，第 1 498 页)或马达拉(Maddala，1983，第 31 页)]，v_t 具有方差，它通过：

$$\bar{\sigma}_t^2=\frac{\bar{p}_t(1-\bar{p}_t)}{N_t[F'(F^{-1}(\bar{p}_t))]^2} \tag{14.26}$$

一致地估计出。伯克森最小卡方估计量(**Berkson's minimum chi-square estimator**) $\hat{\boldsymbol{\beta}}_{MC}$ 是对加权残差和 $\sum_{t=1}^{T}(F^{-1}(\bar{p}_t)-\mathbf{x}_t'\boldsymbol{\beta})/\bar{\sigma}_t^2$ 求关于 $\boldsymbol{\beta}$ 的极小值。这很容易通过 $F^{-1}(\bar{p}_t)/\bar{\sigma}_t$ 对 $\mathbf{x}_t/\bar{\sigma}_t$ 的 OLS 回归计算出来。

这种估计量实施起来简单，原因在于它只需要 OLS 程序包。不过，它是完全有效的，因为可以证明，它与将每个观测值分开处理而不是将观测值分组成含有共同回归元值 $\mathbf{x}_t$ 单元的 MLE 具有相同的渐近分布。对于 logit 模型来说，由于 $F^{-1}(\bar{p}_t)=\ln(\bar{p}_t/(1-\bar{p}_t))$ 且 $\bar{\sigma}_t^2=1/[N_t\bar{p}_t(1-\bar{p}_t)]$，所以该估计量尤其简单。

最小卡方估计量的一个优点是，它计算简单方便，尽管计算机运算能力的不断进步使得这一点不再重要。分组经济数据极少存在，使得每组回归元的唯一值内拥有许多观测值，除非回归元是少数几个指示变量。然而，该方法会提供加总的见解，现在就考察这个专题。

14.6.2 *含有加总数据的估计*

加总数据(**data aggregation**)的经济计量学例子，包括工作人员的比例数据以及住在不同地区乘公交车往返的那些通勤人员的比例数据，这可借助于某地区人员平均特征的数据加以解释。

举一个例子，假定 $\bar{p}_t$ 等于地区 t 的失业率，而 $\mathbf{x}_t$ 等于地区 t 的受教育平均水平。一种可能模型是把 $\bar{p}_t$ 对 $\mathbf{x}_t$ 进行 LS 回归。因为当 $0<\bar{p}_t<1$ 时，许多研究要变

换因变量使其成为无界的,故可估计模型:

$$\ln\left(\frac{\bar{p}_t}{1-\bar{p}_t}\right)=\mathbf{x}_t'\boldsymbol{\beta}+u_t \tag{14.27}$$

其中,u_t 表示误差。

这个模型看起来类似于,当 $F^{-1}(\bar{p}_t)=\ln(\bar{p}_t/(1-\bar{p}_t))$ 时,logit 模型的最小卡方估计量。然而,它却不是,因为只有第 t 个单元(cell)中的所有回归元都取同一值时,伯克森估计量才是适宜的。相反,这里的回归元可取不同值,因为地区 t 的不同人员将具有各不相同的受教育水平。

为了理解回归元存在单元内异质性(**with-cell heterogeneity**)时加总的后果,假定个体水平模型是满足

$$y_i^*=\mathbf{x}_i'\boldsymbol{\beta}+u_i$$
$$u_i\sim\mathcal{N}[0,\ 1]$$

的指标模型(参见 14.4.1 节)。我们选择以正态误差情况开始研究,这对应于 probit 模型而不是 logit 模型,因为可能获得解析结果。对于单元 t 中的个体来说,将异质性建模成:

$$\mathbf{x}_i\sim\mathcal{N}[\boldsymbol{\mu}_t,\ \boldsymbol{\Sigma}_t]$$

现实中允许对不同单元出现变异,而且一种新的复杂问题是 $\boldsymbol{\Sigma}_t\neq\mathbf{0}$,所以存在单元内异质性,于是,在地区 t,以 $\boldsymbol{\beta}$、$\boldsymbol{\mu}_t$ 以及 $\boldsymbol{\Sigma}_t$ 为条件,有:

$$\begin{aligned}\Pr[y_i=1]&=\Pr[(\mathbf{x}_i'\boldsymbol{\beta}+u_i>0]\\&=\Pr\left[\frac{\mathbf{x}_i'\boldsymbol{\beta}+u_i-\mu_t'\boldsymbol{\beta}}{\sqrt{1+\boldsymbol{\beta}'\boldsymbol{\Sigma}_t\boldsymbol{\beta}}}>\frac{-\mu_t'\boldsymbol{\beta}}{\sqrt{1+\boldsymbol{\beta}'\boldsymbol{\Sigma}_t\boldsymbol{\beta}}}\right]\\&=\Phi\left(\frac{\mu_t'\boldsymbol{\beta}}{\sqrt{1+\boldsymbol{\beta}'\boldsymbol{\Sigma}_t\boldsymbol{\beta}}}\right)\end{aligned}$$

其中,运用了已知前面假设时的 $\mathbf{x}_i'\boldsymbol{\beta}+u_i\sim\mathcal{N}[\mu_t'\boldsymbol{\beta},\ (1+\boldsymbol{\beta}'\boldsymbol{\Sigma}_t\boldsymbol{\beta})]$,然后减去均值,再用标准差去除,由此变换成标准正态变量。

已知式(14.24)时,通过类似推理得出式(14.25),基本的个体层次二值选择参数 $\boldsymbol{\beta}$ 能够通过回归:

$$\Phi^{-1}(\bar{y}_t)=\frac{\bar{\mathbf{x}}_t'\boldsymbol{\beta}}{\sqrt{1+\boldsymbol{\beta}'\mathbf{S}_t\boldsymbol{\beta}}}+w_t \tag{14.28}$$

中 $\boldsymbol{\beta}$ 的非线性 LS 估计来得到一致估计,其中,$\bar{y}_t$ 与 $\bar{\mathbf{x}}_t$ 均表示单元 t 平均值,而 $\mathbf{S}_t$ 表示单元 t 中 $\mathbf{x}_i$ 的样本方差。伯克森最小卡方估计却是把 $\Phi^{-1}(\bar{y}_t)$ 对 $\bar{\mathbf{x}}_t$ 进行回归,并且关于 $\boldsymbol{\beta}$ 是非一致的,除非 $\boldsymbol{\Sigma}_t=\mathbf{0}$。

14.6.3 讨论

加总问题在非线性模型中表现得更加复杂。若最初个体水平模型是线性模型 $y_i=\mathbf{x}_i'\boldsymbol{\beta}+u_i$,在第 t 个单元中满足 $\mathbf{x}_i\sim\mathcal{N}[\boldsymbol{\mu}_t,\ \boldsymbol{\Sigma}_t]$,那么相对应的 $\bar{y}_t$ 对 $\bar{\mathbf{x}}_t$ 线性回归会产生 $\boldsymbol{\beta}$ 的一致估计值。就非线性模型而言,类似地,加总会产生个体水平参数

的非一致估计,除非进行调整,使得式(14.28)成立。进一步地,归功于麦克法登和里德(McFadden and Reid, 1975)的14.6.2节中例子是与众不同的,因为非线性模型加总却导致了容易处理的结果。这个例子曾由卡梅伦(Cameron, 1990)做出相当详细的讨论,他在非线性模型加总(**aggregation in nonlinear models**)的较广泛背景下加以考察。

加总的活跃领域——离散选择,通常是多项式选择方面关于品牌商品市场份额的市场营销文献。艾伦比和罗斯(Allenby and Ross, 1991)曾经阐述拟合加总logit模型偏倚可能不是很大的例子。更为重要的是,最近的计算进展使得含有加总数据的个体层次参数的估计成为可能,即使加总会没有产生闭形式解。例如,参见贝里和内沃(Berry and Nevo, 2001),他们在性质上类似于15.7节中随机参数logit模型对模型进行估计。

最后,注意到,在含有加总比例数据的许多应用中,诸如地区失业率,不存在对个体层次参数进行估计的愿望。唯一目标是,对因变量$\bar{p}_t$位于0与1之间做出一个合理模型。于是,线性模型(14.27)或许是优秀的。式(14.27)的误差将不再具有式(14.26)给出的方差。不过,它将仍是异方差的,故统计推断应建立在怀特异方差稳健标准误差的基础上。

14.7 半参数估计

二值结果模型或许是半参数回归的重要例子。大多数经济计量学研究假定单指标形式$F(\mathbf{x}_i'\boldsymbol{\beta})$,其中,关于$F$的函数形式是没有设定的。目标是获得$\boldsymbol{\beta}$的如下估计值:关于$\boldsymbol{\beta}$是一致的、理想上$\sqrt{N}$一致的且渐近正态的,而$F(\cdot)$被认为是冗余函数。人们能应用9.7.4节的单指标模型半参数估计量。另一些估计量探讨了指标函数模型的二值结果的解释。此外,达到半参数有效界的半参数ML估计是可能的,这一方法很少需要额外假设,因为很明显,分布是贝努利分布,而且仅有$F(\mathbf{x}_i'\boldsymbol{\beta})$是未知的。

14.7.1 *半参数条件均值估计*

估计问题通常是因变量y_i取值0或1,满足条件均值:

$$\mathrm{E}[y_i|\mathbf{x}_i]=m(\mathbf{x}_i)$$

其中,$m(\cdot)$表示未知的。注意到,$m(\mathbf{x}_i)$同样等于$y_i=1$的条件概率。

不管因变量的二值性质怎样,可应用9.4节至9.6节的非参数回归方法。这很容易从图14.1中看出,二值变量y对纯量回归元x的散点图,作为y对x的核回归的一个自然备选者。若暂且不谈异常情况,例如当使用较高阶核时,在此情况下拟合值能取负值,则拟合值将位于0与1之间。

在许多微观经济计量学应用中,就发挥良好作用的非参数方法来说,$\mathbf{x}$具有太高维度(维数祸根)。部分设定$m(\cdot)$的半参数回归模型已由9.7节给出。可加模型在条件应用中相当流行。在经济计量学中,反而运用单指标模型,因为受欢迎的

起点是 14.4.1 节指标函数模型。当潜变量 $y^* = \mathbf{x}'\boldsymbol{\beta} + u$ 时，就得到单指标模型（**single-index model**）。因而，我们有：

$$\mathrm{E}[y_i \mid \mathbf{x}_i] = F(\mathbf{x}_i'\boldsymbol{\beta})$$

其中，遵循本章的记号，用 $F(\cdot)$ 而不是 $g(\cdot)$ 表示未知函数。

由 9.7.4 节知，$\boldsymbol{\beta}$ 是唯一可识别的，至多相差位置与标度。由 14.4.1 节知，这很明显，其中，指标模型中的误差 u 被正规化成具有 0 均值（位置），且其方差需要加以设定（标度）。此处，对 u 没有施加约束，因此 β 不是完全可识别的，但斜率系数的比率是可识别的。参见曼斯基（Manski, 1988b）对二值选择模型识别的详细分析。

$\boldsymbol{\beta}$ 的一致渐近正态估计值能通过平均求导数估计，或通过半参数最小二乘法来获得（参见 9.7.4 节）。不过，针对二值结果特有的一些可供选择估计量更经常被使用。

14.7.2 *最大得分估计*

二值结果的半参数估计量常常建立在关于二值结果的指标函数模型 $y^* = \mathbf{x}'\boldsymbol{\beta} + u$ 基础上。在这种情况下，将模型写成：

$$y_i = \mathbf{1}(x_i'\boldsymbol{\beta} + u_i > 0)$$

会很方便，其中，当事件 A 发生时，有 $\mathbf{1}(A) = 1$。

曼斯基（Manski, 1975）发现，由于 u_i 是未知的，令 $u_i = 0$，y_i 的预测值是 $\mathbf{1}(\mathbf{x}_i'\boldsymbol{\beta} > 0)$，在此情况下，正确预测次数的得分为：

$$S_N(\boldsymbol{\beta}) = \sum_{i=1}^{N} \{y_i \mathbf{1}(\mathbf{x}_i'\boldsymbol{\beta} > 0) + (1 - y_i)\mathbf{1}(\mathbf{x}_i'\boldsymbol{\beta} \leqslant 0)\} \tag{14.29}$$

因为当 $y_i = 1$ 且 $\mathbf{1}(\mathbf{x}_i'\boldsymbol{\beta} > 0)$ 时，或当 $y_i = 0$ 且 $\mathbf{1}(\mathbf{x}_i'\boldsymbol{\beta} \leqslant 0)$ 时，都会得到正确预测。曼斯基的最大得分估计量（**maximum score estimator**）是求 $S_N(\boldsymbol{\beta})$ 极大值的解。这是一个非标准问题，因为 $\mathbf{1}(\mathbf{x}_i'\boldsymbol{\beta} > 0)$ 在 $\boldsymbol{\beta}$ 处不可微。曼斯基（Manski, 1975, 1985）已经建立了一致性假设，或等价地有，$\mathrm{Median}[u_i \mid \mathbf{x}_i] = 0$。然后，可以证明，$N^{1/3}(\hat{\boldsymbol{\beta}}_{\mathrm{ms}} - \boldsymbol{\beta})$ 服从非正态极限分布，尽管推断可利用自助法来执行[曼斯基和汤普森（Manski and Thompson, 1986）]。

曼斯基估计量可被看成最小绝对偏差估计量。由 4.6.2 节知，LAD（最小绝对偏差）估计量是求 y_i 与 $\mathrm{Median}[y_i \mid \mathbf{x}_i]$ 之间绝对差之和的最小值。这种不熟悉的估计量，在性质上类似于 LS 估计量，LS 估计量是求 y_i 与 $\mathrm{E}[y_i \mid \mathbf{x}_i]$ 之间绝对差之和的最小值。为了实施 LAD，此处需要获得 $\mathrm{Median}[y_i \mid \mathbf{x}_i]$。当 $\mathrm{Median}[u_i \mid \mathbf{x}_i] = 0$ 时，$\mathrm{Median}[y_i^* \mid \mathbf{x}_i] = \mathbf{x}_i'\boldsymbol{\beta}$，所以 $\mathrm{Median}[y_i \mid \mathbf{x}_i] = \mathbf{1}(\mathbf{x}'\boldsymbol{\beta} > 0)$。因此，二值结果模型 LAD 估计量（**binary outcome model LAD estimator**）是求

$$Q_N(\boldsymbol{\beta}) = \sum_{i=1}^{N} | y_i - \mathbf{1}(\mathbf{x}_i'\boldsymbol{\beta} > 0) | \tag{14.30}$$

的极小值。由习题 14.4 知，$Q_N(\boldsymbol{\beta}) = N - S_N(\boldsymbol{\beta})$，故最大得分估计量等于 LAD 估

计量。关于最大得分估计量作为 LAD 估计量的其他一些解释，可参见曼斯基(Manski, 1985,第 320 页)。

由式(14.29)给出的最大得分估计量，其目标函数 $S_N(\boldsymbol{\beta})$不是可微的。它能重新写成：

$$S_N(\boldsymbol{\beta}) = \sum_{i=1}^{N}(2y_i - 1)\mathbf{1}(\mathbf{x}_i'\boldsymbol{\beta} > 0) + N - \sum_{i=1}^{N} y_i$$

参见习题 14.4。第二个求和可被忽略掉，因为它不涉及 $\boldsymbol{\beta}$。

具有可微目标函数的估计量是霍罗威茨(Horowitz, 1992)的光滑最大得分估计量(**smooth maximum score estimator**)，它是求

$$Q_N^S(\boldsymbol{\beta}) = \sum_{i=1}^{N}(2y_i - 1)K(\mathbf{x}_i'\boldsymbol{\beta}/h_N)$$

的极大值，其中，$K(\mathbf{x}'\boldsymbol{\beta}/h_N)$表示 $\mathbf{1}(\mathbf{x}'\boldsymbol{\beta}>0)$的光滑形式。由于对 $\mathbf{x}'\boldsymbol{\beta}$的负值来说，$\mathbf{1}(\mathbf{x}'\boldsymbol{\beta}>0)$等于 0，并且对 $\mathbf{x}'\boldsymbol{\beta}$的正值来说，$\mathbf{1}(\mathbf{x}'\boldsymbol{\beta}>0)=1$，所以选取 $K(\cdot)$为满足 $K(0)=0.5$ 的 cdf，同时选取 h_N 为很小的，这样做是很自然的。光滑使得该估计量的计算得以简化，但分析却是错综复杂的，因为要求当 $N\to\infty$时，h_N 必须以适当速率 $h_N\to 0$。此估计量以接近于$\sqrt{N}$的速率收敛。对于详细内容，参见霍罗威茨(Horowitz, 2002)，他曾经阐述有限样本中允许检验含有较好水平性质的自助法(使检验具有较好水平性质成为可能的自助法)。

可将 LAD 估计量推广到删失回归模型(参见 16.9.2 节)。

14.7.3 *最大秩相关估计量*

以满足 $\mathrm{E}[y_i|\mathbf{x}_i]=F(\mathbf{x}_i'\boldsymbol{\beta})$的单指标模型开始研究。若 $F(\mathbf{x}_i'\boldsymbol{\beta})$关于 $\mathbf{x}_i'\boldsymbol{\beta}$是单调递增的，当 $\mathbf{x}_i'\boldsymbol{\beta}>\mathbf{x}_j'\boldsymbol{\beta}$，则 $\mathrm{E}[y_i|\mathbf{x}_i]>\mathrm{E}[y_j|\mathbf{x}_j]$。因而，虽然不能保证下述情况，但可能会出现当 $\mathbf{x}_i'\boldsymbol{\beta}>\mathbf{x}_j'\boldsymbol{\beta}$时，观测值 $y_i>y_j$。这就建议了，当 $\mathbf{x}_i'\boldsymbol{\beta}>\mathbf{x}_j'\boldsymbol{\beta}$时，选取 $\boldsymbol{\beta}$ 确保高频数 $y_i>y_j$。

哈恩(Han, 1987)的最大秩相关(**maximum rank correlation, MRC**)估计量选择 $\boldsymbol{\beta}$，使：

$$Q_N^{\mathrm{MRC}}(\boldsymbol{\beta}) = \sum_{i=1}^{N}\sum_{\substack{j=1\\ j<i}}^{N}\mathbf{1}(y_i > y_j)\mathbf{1}(\mathbf{x}_i'\boldsymbol{\beta} > \mathbf{x}_j'\boldsymbol{\beta}) + \mathbf{1}(y_i < y_j)\mathbf{1}(\mathbf{x}_i'\boldsymbol{\beta} < \mathbf{x}_j'\boldsymbol{\beta})$$

若当 $\mathbf{x}_i'\boldsymbol{\beta}>\mathbf{x}_j'\boldsymbol{\beta}$ 时 $y_i>y_j$，或若当 $\mathbf{x}_i'\boldsymbol{\beta}<\mathbf{x}_j'\boldsymbol{\beta}$时 $y_i<y_j$，则此和式中第 ij 项等于 0，而若存在符号反向的情况，以致当 $\mathbf{x}_i'\boldsymbol{\beta}>\mathbf{x}_j'\boldsymbol{\beta}$时 $y_i<y_j$，或者 $\mathbf{x}_i'\boldsymbol{\beta}<\mathbf{x}_j'\boldsymbol{\beta}$时 $y_i>y_j$，则第 ij 项等于 1。将该估计量称为最大秩相关估计量，因为 $Q_N^{\mathrm{MRC}}(\boldsymbol{\beta})$是 y_i 与 $\mathbf{x}_i'\boldsymbol{\beta}$之间肯德尔秩相关系数的倍数。

这个估计量是$\sqrt{N}$一致的且渐近正态的[参见舍曼(Sherman, 1993)]。

14.7.4 *半参数 ML 估计*

对于二值选择数据来说，给定独立观测值时，似然函数显然是由式(14.4)给出的形式。唯一的复杂情况是 $F(\cdot)$为未知的。克莱因和斯帕迪(Klein and Spady,

1993)已经提出半参数 MLE(**semiparametric MLE**),它是对

$$\mathcal{L}_N(\boldsymbol{\beta}) = \sum_{i=1}^{N} \{y_i \ln \hat{F}(\mathbf{x}_i'\boldsymbol{\beta}) + (1-y_i)\ln(1-\hat{F}(\mathbf{x}_i'\boldsymbol{\beta}))\}$$

求极大值,其中,$\hat{F}(\mathbf{x}_i'\boldsymbol{\beta})$表示 $F(\mathbf{x}_i'\boldsymbol{\beta})$的非参数估计值。

这一估计量,在思想上类似于 9.7.4 节详述的市村(Ichimura, 1993)WSLS 估计量,并且在给定 $\hat{F}$ 时计算$\hat{\boldsymbol{\beta}}$ 与给定$\hat{\boldsymbol{\beta}}$ 时计算 $\hat{F}$ 之间,迭代计算时会出现类似问题。已知 ML 一阶条件(14.5),半参数 MLE 还能被计算成方程

$$\sum_{i=1}^{N} \frac{\hat{F}'(\mathbf{x}_i'\boldsymbol{\beta})}{\hat{F}(\mathbf{x}_i'\boldsymbol{\beta})(1-\hat{F}(\mathbf{x}_i'\boldsymbol{\beta}))}(y_i - \hat{F}(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{x}_i = \mathbf{0}$$

的解,这与含有权数 $w_i=\hat{F}_i'/[\hat{F}_i(1-\hat{F}_i)]$ 的 WSLS 估计量(**WSLS estimator**)的那些情况一样。

克莱因和斯帕迪估计量的吸引人之处是,在它达到半参数有效界时是完全有效的。不过,计算极为困难。详细内容参见 9.7.4 节,其中类似计算问题曾对市村的 WSLS 估计量讨论过,并参见克莱因和斯帕迪(Klein and Spady, 1993),以及帕甘和乌拉(Pagan and Ullah, 1999,第 283～285 页)。

14.7.5 半参数估计量的比较

经济计量学家关注单指标模型,而且甚至于对二值结果模型来说,存在大量可利用的半参数估计量。这些估计量中的任一个都不是特别简单易行的。目标函数具有多重最优值且不是光滑的。例如,霍罗威茨(Horowitz, 1992)运用光滑最大得分估计量的模拟退火,而多西尔和迈耶(Dorsey and Mayer, 1995)使用遗传算法来获得最大得分估计量。

对系数进行解释同样是困难的。例如,用于钓鱼方式数据的最大得分估计量会得出 0.776 截距估计值,而−0.631 斜率估计值(其自助法估计的标准误差为 0.103),但这些系数都不可直接与表 14.2 给出的那些值进行比较。实际上,由于参数斜率估计值至多差一个标度都是恰好识别的,所以如果几个系数都包含在回归中,且系数估计值都可以与参考变量的那些值进行比较,那么半参数估计值是相当有用的。

在不需要使用光滑系数诸如带宽选择这一引人注目的性质的半参数估计量之间,最大得分估计量与最大秩相关估计量均与众不同。这两个估计量中,后者是$\sqrt{N}$一致的。

在最近的研究工作中,布伦德尔和鲍威尔(Blundell and Powell, 2004)曾经提出了含有内生回归元(**endogenous regressors**)的半参数估计。

14.8 第 I 类极值的 logit 推导

源自 ARUM 的 14.4.2 节 logit 模型的推导使用了下述条件结果的知识:独立的第 I 类极值随机变量的差 $\varepsilon_0-\varepsilon_1$ 是逻辑斯蒂分布。为了完整起见,我们提供建

立在 ε_0 与 ε_1 分布基础上的直接推导。

将式(14.20)的第二行重新写出，得到：

$$\begin{aligned}\Pr[y=1] &= \Pr[\varepsilon_0 < \varepsilon_1 + V_1 - V_0] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\varepsilon_1+V_1-V_0} f(\varepsilon_0, \varepsilon_1)d\varepsilon_0 d\varepsilon_1 \\ &= \int_{-\infty}^{\infty} f(\varepsilon_1)\left\{\int_{-\infty}^{\varepsilon_1+V_1-V_0} f(\varepsilon_0)d\varepsilon_0\right\}d\varepsilon_1\end{aligned} \tag{14.31}$$

其中，最后一行 ε_0 与 ε_1 均被假定成独立的。通过将 $f(\varepsilon_0)$限定成第I类极值密度，式(14.31)变成：

$$\begin{aligned}\Pr[y=1] &= \int_{-\infty}^{\infty} f(\varepsilon_1)\left\{\int_{-\infty}^{\varepsilon_1+V_1-V_0} e^{-\varepsilon_0}\exp(-e^{-\varepsilon_0})d\varepsilon_0\right\}d\varepsilon_1 \\ &= \int_{-\infty}^{\infty} f(\varepsilon_1)[\exp(-e^{-\varepsilon_0})]_{-\infty}^{\varepsilon_1+V_1-V_0} d\varepsilon_1 \\ &= \int_{-\infty}^{\infty} f(\varepsilon_1)\exp(-e^{-(\varepsilon_1+V_1-V_0)})d\varepsilon_1\end{aligned} \tag{14.32}$$

利用式(14.32)关于 ε_1 的极值密度，得到：

$$\begin{aligned}\Pr[y=1] &= \int_{-\infty}^{\infty} e^{-\varepsilon_1}\exp(-e^{-\varepsilon_1})\exp(-e^{-(\varepsilon_1+V_1-V_0)})d\varepsilon_1 \\ &= \int_{-\infty}^{\infty} e^{-\varepsilon_1}\{\exp(-e^{-\varepsilon_1} - e^{-(\varepsilon_1+V_1-V_0)})\}d\varepsilon_1 \\ &= \int_{-\infty}^{\infty} e^{-\varepsilon_1}\{\exp(-e^{-\varepsilon_1} - e^{-\varepsilon_1}e^{-(V_1-V_0)})\}d\varepsilon_1 \\ &= \int_{-\infty}^{\infty} e^{-\varepsilon_1}\exp\{-e^{-\varepsilon_1}(1+e^{-(V_1-V_0)})\}d\varepsilon_1\end{aligned} \tag{14.33}$$

由于 $\int_{-\infty}^{\infty} ae^{-\varepsilon}\exp(-ae^{-\varepsilon})d\varepsilon = 1$，可得$\int_{-\infty}^{\infty} e^{-\varepsilon}\exp(-ae^{-\varepsilon})d\varepsilon = 1/a$。利用式(14.33)及 $a = 1+e^{-(V_1-V_0)}$，得到：

$$\begin{aligned}\Pr[y=1] &= (1+e^{-(V_1-V_0)})^{-1} \\ &= e^{V_1}/(e^{V_0}+e^{V_1}) \\ &= e^{V_1-V_0}/(1+e^{V_1-V_0})\end{aligned} \tag{14.34}$$

若令 $V_1 - V_0 = \mathbf{x}'\boldsymbol{\beta}$，则得到 logit 模型。

14.9 应用研究

大多数软件包都提供 probit 与 logit 模型估计量。对于应用者来说，主要抉择是运用哪一个模型。在实际应用中，除非大部分结果都是 0 或大部分结果都是 1，否则从这两个模型获得的预测边际效应差异很小。

尽管 Lindep 可实施曼斯基以及克莱因和斯帕迪估计量，但通常半参数估计要求运用诸如 GAUSS 语言进行特殊编程。

14.10 文献注释

logit 与 probit 模型是广泛运用且相对简单的非线性回归模型，它们出现在许多标准教科书中，例如格林(Greene，2003)的书。由雨宫(Amemiya，1981)与麦克法登(McFadden，1984)撰写的综述包含了所有基本结果。马达拉(Maddala，1983)与雨宫(Amemiya，1985)的书提供了更为详细的内容。在应用方面，特雷恩(Train，1986)、本·阿基瓦和莱尔曼(Ben-Akiva and Lerman，1985)的书是特别好的。这些参考书既涵盖二值结果，又涵盖多项式结果。

14.3 为了画出剂量死亡率曲线(dosage-mortality curves)，布利斯(Bliss，1934)提供 probit 变换。伯克森(Berkson，1951)则推动了最简单 logit 模型的广泛运用。

14.4 潜变量模型在心理测验文献中尤其流行。

14.5 雨宫(Amemiya，1985，9.5 节)提供二值结果模型的基于选择抽样的一个优秀综述。也可参见 24.4 节。

14.6 卡梅伦(Cameron，1990)考察二值结果模型中的加总问题，同时对凯利吉安(Kelijian，1980)与斯托克(Stoker，1984)关于利用加总数据的非线性模型中个体水平参数可估计性的一般性结果加以归纳总结。

14.7 曼斯基(Manski，1975)的最大得分估计量是半参数回归的早期一个重要例子。关于二值结果模型的半参数方法已由李明宰(M-J. Lee，1996)、霍罗威茨(Horowitz，1997)以及帕甘和乌拉(Pagan and Ullah，1999)撰写的书涵盖。后者文献中涵盖了许多方法。

习 题

14-1 考察由 $y_i^*=\mathbf{x}_i'\boldsymbol{\beta}+\varepsilon_i$ 建立的潜变量模型，其中 $\varepsilon_i\sim\mathcal{N}[0,1]$。假定只有当 $y_i^*<U_i$ 时，观测到 $y_i=1$，而只有当 $y_i^*\geq U_i$ 时，观测到 $y_i=0$，就每个个体而言，上极限 U_i 是已知常值，而对不同个体来说可能是不同的。

(a) 求 $\Pr[y=1|\mathbf{x}_i]$。[提示：注意到，这既由于 U_i 的存在而不同于标准情况，又因为当 $y_i^*<U_i$ 时有 $y_i=1$ 而要转变一些等式。]

(b) 请提供一致估计 $\boldsymbol{\beta}$ 的估计方法的细节。

(c) 假定估计这一模型，并求第三个回归元 x_{3i} 具有估计系数 $\hat{\beta}_3=0.2$。给出 $\hat{\beta}_3$ 有意义的解释。

14-2 考察满足 $\Pr[y=1|x_1,x_2]=\Lambda(\beta_0+\beta_1x_{1i}+\beta_2x_{2i})$ 的 logit 模型，其中，$\Lambda(z)=e^z/(1+e^z)$。

(a) 以推广形式写出似然得分以及信息矩阵。

(b) 利用这些推导沃尔德检验以及 LM 得分检验 $H_0:\beta_2=0$。

(c) 请解释你如何在计算上实施检验。

(d) logit 模型在什么意义下内在地成为异方差？

14-3 假定对离散选择模型使用指标公式，但认为潜变量是严格正的。这通过假定潜变量 y^* 具有指数密度而得以适应，其参数为 γ，因此，密度 $f(y^*)$ 成为 $f(y^*)=\gamma^{-1}\exp(-y^*/\gamma)$，满足 $\gamma=\exp(\mathbf{x}'\boldsymbol{\beta})$。当 $y^*>\mathbf{z}'\boldsymbol{\alpha}$ 时，才能观测到 $y=1$；而当 $y^*\leqslant\mathbf{z}'\boldsymbol{\alpha}$ 时，才能观测到 $y=0$。

(a) 给出观测数据的对数似然函数。

(b) 当 x_{ji} 变动一个单位时，对 $\Pr[y_i=1]$ 的效应是多少？

(c) 假定当 $y^*>\exp(\mathbf{z}'\boldsymbol{\alpha})$ 且 $\mathbf{x}=\mathbf{z}$ 时，$y=1$。在识别 $\boldsymbol{\alpha}$ 以及/或者 $\boldsymbol{\beta}$ 时，你会看到什么问题？请解释你的答案。

14-4 考察含有式(14.29)给出的目标函数 $S_N(\boldsymbol{\beta})$ 以及(14.30)给出的 $Q_N(\boldsymbol{\beta})$ 的最大得分估计量。

(a) 证明 $S_N(\boldsymbol{\beta})=\sum_i[\mathbf{1}(y_i=1)\times\mathbf{1}(\mathbf{x}_i'\boldsymbol{\beta}>0)+\mathbf{1}(y_i=0)\times\mathbf{1}(\mathbf{x}_i'\boldsymbol{\beta}\leqslant0)]$。

(b) 证明 $Q_N(\boldsymbol{\beta})=\sum_i[\mathbf{1}(y_i=1)\times\mathbf{1}(\mathbf{x}_i'\boldsymbol{\beta}\leqslant0)+\mathbf{1}(y_i=0)\times\mathbf{1}(\mathbf{x}_i'\boldsymbol{\beta}>0)]$。

(c) 利用 $\mathbf{1}(y_i=1)=1-\mathbf{1}(y_i=0)$，证明 $Q_N(\boldsymbol{\beta})=N-S_N(\boldsymbol{\beta})$ 。

(d) 利用 $\mathbf{1}(\mathbf{x}_i'\boldsymbol{\beta}\leqslant0)=1-\mathbf{1}(\mathbf{x}_i'\boldsymbol{\beta}>0)$，证明式(14.29)能重新写成 $S_N(\boldsymbol{\beta})=\sum_i(2y_i=1)\mathbf{1}(\mathbf{x}_i'\boldsymbol{\beta}>0)+N-\sum_i y_i$。

14-5 运用 16.6 节的健康消费数据。模型是 DMED 的 probit 模型。DMED 表示正的健康消费的指示变量，为了简单起见，仅对应于单一回归元 NDISEASE，即慢性病数量。

(a) 求斜率参数的 OLS 估计。

(b) 求斜率参数的 probit 估计。

(c) 已知(b)部分，以两种方式求出慢性病的边际效应：样本的平均值以及 NDISEASE 样本平均值处的估计。

(d) 求斜率参数的 logit 估计。

(e) 已知(d)部分，以三种方式求出慢性病的边际效应：样本的平均值、NDISEASE 样本平均值处的估计以及在 $\Lambda(\mathbf{x}'\boldsymbol{\beta})=\bar{y}$ 处的估计。

(f) 对于 logit 模型来说，当 NDISEASE 变化时，计算优势比的比例变化。

14-6 继续习题 14.5 的分析。

(a) 根据 NDISEASE 的统计显著性，比较三个二值模型。

(b) 根据估计的边际效应，比较三个二值模型。

(c) 根据预测概率，比较三个二值模型。

(d) 根据对数似然，比较 logit 二值模型与 probit 二值模型。

15
多项式模型

15.1 引 论

前面一章已经考察了离散变量取两个可能值之一的一些模型。本章考察具有几种可能结果的模型，其中几种可能结果通常是互不相交的。一些例子包括：上下班往返通勤采用的不同方式（乘公交车、小车或步行）、各种健康保险类型（一次一付医疗费、管理医疗或没有参加）、各种不同就业状况（全日制、兼职或没有工作）、娱乐地点的选择、职业选择以及产品选择。

正如二值数据必服从贝努利分布或二项式分布一样，由于数据必服从多项式分布，所以原则上统计推断相对简单直接。因为数据显然是服从多项式分布的，最常见的估计是通过极大似然来完成。可是，对于某些复杂情况来说，反而用基于矩的估计。

类似于二值情况 probit 与 logit 之间的差别，因多项式分布概率有各种函数形式，故产生了各种不同的多项式模型。在这些模型之间，同样可区分，给定个体时有些回归元会随选项不同而变化的模型，以及有些回归元随选项不同而为常值的模型。例如，在运输方式选择中，一些回归元诸如旅行次数或成本将会随选项不同而变化，而其他一些回归元譬如年龄却是不随选项而变化的。

一种最简单的多项式模型，即条件 logit 模型或多项式的 logit 模型，运用起来相当简单易行，但在实际应用中却被认为其约束性太强，尤其是当多项式结果数据源自个体选项时。对于无序结果来说，稍欠约束的模型能利用随机效用模型来得到。在此模型中，具有最高效用的选项被选上，其中，来自每一个选项的效用都是确定性成分之和。对随机成分的各种不同设定导致了选择概率的各种函数形式，从而产生各种不同的多项式模型。在一些应用中，对决策过程施加某种结构时，诸如选项的自然顺序或者决策次序，就会出现其他模型。人们会在实际中应用许多不同的多项式模型。

15.2 节运用例子阐述本章将要讨论的问题。15.3 节给出多项式模型的一般性结果。条件与多项式 logit 模型由 15.4 节讨论。可加随机效用模型放在 15.5 节阐述。嵌入式 logit、随机参数 logit 以及多项式 probit 模型将是 15.6 节至 15.8 节的主题。有序与时序模型将在 15.9 节详述。拥有多于一个离散变量的多变量

模型在 15.10 节加以阐述。半参数估计量则在 15.11 节简略讨论。

15.2 例子:钓鱼方式的选择

本节阐述多项式,即最简单的无序多项式模型,并在 15.4 节详述允许回归元随着选项而变化的一些变形。强调内容放在对估计模型的解释上。与通常单个条件均值的影响相比,回归元变化的边际效应更为复杂。对于多项式数据来说,每一个结果都存在各自对概率的边际效应,同时由于这些概率之和为 1,所以其边际效应之和为 0。

对钓鱼方式的选择就是一个应用。因变量 y 取值 1、2、3 或 4,它们分别表示选择岸边、码头、私家船以及租船这四种相互排斥的钓鱼方式。无序多项式模型,诸如多项式 logit,适合于钓鱼方式的选择,因为结果变量不存在明显的排序关系。回归元是个体收入、价格以及捕获率,其中,个体收入并不随钓鱼方式而变化,价格与捕获率则会随钓鱼方式以及不同个体而变化。

1 182 个人员的样本来自汤姆森和克鲁克(Thomson and Crooke, 1991)实施的调查,并由赫里格斯和克林(Herriges and Kling, 1999)加以分析的研究。表 15.1 对这些数据进行了概括,给出选择每一种方式人员子样本的平均值以及回归元的整个样本平均值。

表 15.1 钓鱼方式多项式选择:数据概括

	子样本均值				
解释变量	$y=1$ 岸边	$y=2$ 码头	$y=3$ 私家船	$y=4$ 租船	所有 y
收入(每月 1 000 美元)	4.052	3.387	4.654	3.881	4.099
岸边价格(美元)	36	31	138	121	103
码头价格(美元)	36	31	138	121	103
私家船价格(美元)	98	82	42	45	55
租船价格(美元)	125	110	71	75	84
岸边捕获率	0.28	0.26	0.21	0.25	0.24
码头捕获率	0.22	0.20	0.13	0.16	0.16
私家船捕获率	0.16	0.15	0.65	0.69	0.63
岸边捕获率	0.52	0.50	0.65	0.69	0.63
样本概率	0.113	0.151	0.354	0.382	1.000
观测值	134	178	418	452	1 182

15.2.1 *条件 logit: 选项变化回归元*

首先,考察价格与捕获率的作用,回归元会随着选项不同而变化,只是对这些数据而言,岸边和码头钓鱼的价格都是相同的。

沿着表 15.1 的列往下看,可以发现,人们趋向于最便宜的钓鱼方式。例如,与选择其他钓鱼方式的平均价格 36 美元、98 美元以及 125 美元相比,对于选择岸边钓鱼的人来说,其平均价格为 36 美元。

更一般地讲，对于选择岸边和码头钓鱼的人来说，与船上钓鱼相比，这两种方式平均更为便宜，而对于船上钓鱼的人来说，与岸边或码头钓鱼相比，平均更为便宜。很明显，尽管租船钓鱼捕获率最大，但在方式选择与捕获率之间的关系含糊不清。

对于随选项而变化的特定选项回归元，诸如价格与捕获率来说，多项式 logit 模型称为条件 logit 模型（参见 15.4.1 节）。第 i 个个体选择第 j 种钓鱼方式的概率为：

$$p_{ij}=\Pr[y_i=j]=\frac{\exp(\beta_P P_{ij}+\beta_C C_{ij})}{\sum_{k=1}^{4}\exp(\beta_P P_{ik}+\beta_C C_{ik})},\quad j=1,\cdots,4$$

其中，P 表示价格，C 表示捕获率，下标 i 表示第 i 个个体，下标 j 或 k 表示选项。该模型是二值 logit 模型的明显推广，而且给出的概率位于 0 与 1 之间且和为 1。其他一些多项式模型则使用 p_{ij} 的不同函数形式。

系数估计，已由表 15.2 的 CL 列给出。对于 CL 模型来说，即使不是所有多项式模型，但对系数符号可直接进行解释。由于 $\beta_P<0$，由 15.4.3 节，可以预期到，一个选项的价格增加会减少选择该选项的概率，从而使得选择其他选项的概率增大。类似地，由于 $\beta_C>0$，所以一个选项的捕获率增加会增大选择该选项的概率，从而使得选择选项的概率减少。

表 15.2 钓鱼方式多项式选择：logit 估计[a]

回归元	类型	系数	模型形式		
			CL	MNL	混合的
价格 (P)	特定的	β_P	−0.021	—	−0.025
捕获率 (C)	特定的	β_{CR}	0.953	—	0.358
截距	不变的	α_1：岸边	—	0.0	0.0
		α_2：码头	—	0.814	0.778
		α_3：私家船	—	0.739	0.527
		α_4：租船	—	1.341	1.694
收入	不变的	β_{I1}：	—	0.0	0.0
		β_{I2}：	—	−0.143	−0.128
		β_{I3}：	—	0.092	0.089
		β_{I4}：	—	−0.032	−0.033
$-\ln L$			−1 311	−1 477	−1 215
伪 R^2			0.162	0.099	0.258

[a] 回归元类型是特定选项（价格与捕获率）或不变选项（收入）。结果是：(1) 岸边；(2) 码头；(3) 私家船；(4) 租船。MLE 估计值是条件 logit(CL)、多项式 logit(MNL)以及混合 logit(混合)模型。MNL 模型与混合模型被正规化为基准岸边类别。除了 β_{I4} 之外，所有估计值在 5%上都是统计显著的。

对回归元变动所产生影响的标准测量是 $N^{-1}\sum_{i=1}^{N}\partial p_{ij}/\partial x_{ikr}$，即对于第 k 种选项来说，当第 r 个回归元增加一个单位时，选择第 j 项概率的平均边际响应，同时对其他选项来说则是不变的。对于 CL 模型而言，这可通过 $N^{-1}\sum_{i=1}^{n}\hat{p}_{ij}(\delta_{ijk}-\hat{p}_{ik})\hat{\beta}_r$ 估计[参见式(15.38)]，其中，$\hat{\beta}$ 表示 β 的估计值，而 $\hat{p}_{ij}$ 表示预测概率，$j=1,\cdots,m$。

对于四种不同模型的两个回归元即价格与捕获率的平均响应，已由表 15.3 给出。该表给出价格上变动 100 个单位（或 100 美元）时，选择概率的效应以及捕获率变动一个单位时的效应。例如，岸边钓鱼价格增加 100 美元时，分别导致岸边钓鱼、码头钓鱼、私家船钓鱼以及租船钓鱼的概率减少 0.272、增大 0.119、增大 0.080 以及增大 0.068。注意到，如人们所料，其概率变化之和为 0。

表 15.3　钓鱼方式选择：条件 logit 模型的边际效应[a]

	价格变化 100 美元				捕获率变化一个单位			
	岸边	码头	私家船	租船	岸边	码头	私家船	租船
Pr[岸边]变化	−0.272	0.119	0.085	0.068	0.126	−0.055	−0.040	−0.032
Pr[码头]变化	0.119	−0.263	0.080	0.064	−0.055	0.122	−0.037	−0.030
Pr[私家船]变化	0.080	0.080	−0.391	0.225	−0.040	−0.037	0.182	−0.105
Pr[租船]变化	0.068	0.064	0.225	−0.357	−0.032	−0.030	−0.105	0.166

[a] 当回归元对其中一个选项发生变化而对其他选项不变时，选取每个选项的概率的平均边际响应。

这些边际效应与概率的计算需要估计之后来进行计算。对于 CL 模型来说，快速计算[1]（back-of-the-envelop calculation）使用了 $\bar{p}_j(\delta_{jk}-\bar{p}_k)\hat{\beta}$，其中，$\bar{p}_j$ 表示样本平均概率。对于岸边钓鱼价格变化 100 美元对岸边钓鱼概率的效应来说，$100\times0.113\times(1-0.113)\times(-0.21)=-0.21$，与表中样本平均值 −0.272 相比，当概率比较接近于 0 或 1 时，这种近似变得缺少合理性。

表 15.3 中的结果与下述观点一致：最大的替代关系是在码头钓鱼与岸边钓鱼、私家船钓鱼与租船之间。具体地讲，对于码头钓鱼来说，价格上升或捕获率下降都会导致用去岸边钓鱼作为替代，反之亦然。对租船钓鱼与私家船钓鱼来说，类似结果仍成立。

倘若平均价格为 86 美元且平均捕获率为 0.30，这些概率变动在回归元上显得非常大。不过，人们可以计算弹性。使用选择概率需要小心慎重，因为概率位于 0 与 1 之间是有界的。当预测概率从 0.01 到 0.02 变动时所产生的弹性大致是预测概率从 0.50 到 0.51 变动时所产生弹性的 50 倍。

15.2.2　*多项式 logit：选项不变回归元*

现在，考察以千美元测算的每月收入的作用。由表 15.1 知，可以看出，当收入提高时，钓鱼方式会依次从码头钓鱼到租船钓鱼再到岸边，而最终到私家船钓鱼，这里在码头钓鱼的人员平均月收入为 3 387 美元，而私家船钓鱼的人员平均月收入为 4 654 美元。

因为收入对选项来说是不变的，所以合适的模型是多项式 logit 模型（将在 15.4.1 节阐述）。这将设置回归元系数随选项而变化，满足：

---

〔1〕是指测试一个假设的粗略计算，它不一定写在信封的背面。通常，该术语比猜想要可信一些，但是不如一个数学定理那样确定。它经常用于数学、物理和某些工程领域。这里将它译成快速计算。——译者注

$$p_{ij}=\Pr[y_i=j]=\frac{\exp(\alpha_j+\beta_{Ij}I_i)}{\sum_{k=1}^{4}\exp(\alpha_k+\beta_{Ik}I_i)}, \quad j=1,\cdots,4$$

其中，I 表示收入。由于约束的概率之和为 1，需要对参数加以正规化。经验结果，令 $\alpha_1=0$ 且 $\beta_{I1}=0$。

参数估计值，已由表 15.2 中的 MNL 列给出。与 CL logit 模型相比，对其系数给出解释就更加困难。特别地，对于 MNL 模型来说，正的回归系数并不意味着，回归元增大会导致那个选项概率的增加。相反，对 MNL 模型的解释与参照或基准类别组有关，此处作为岸边系数的岸边被正规化为 0。与岸边钓鱼相比，较高收入会导致源自码头(由于 $\beta_{I2}=-0.143<0$)或租船(由于 $\beta_{I3}=0.092$)的钓鱼似然，并使私家船钓鱼的似然较大。

对收入变动响应的数量，可用 $N^{-1}\sum_{i=1}^{N}\partial p_{ij}/\partial I_i$ 进行测算，即对个体的边际效应进行平均。就 MNL 模型而言，这通过 $N^{-1}\sum_{i=1}^{N}\hat{p}_{ij}(\hat{\beta}_j-\hat{\bar{\beta}}_i)$ 估计[参见式(15.19)]，其中，$\hat{\boldsymbol{\beta}}_j$ 表示 $\boldsymbol{\beta}_j$ 的估计值，$\bar{\boldsymbol{\beta}}_i=\sum_{l=1}^{m}p_{il}\boldsymbol{\beta}_l$ 表示 $\boldsymbol{\beta}_l$ 的加权概率平均，而 $\hat{p}_{ij}$ 表示预测概率，$j=1,\cdots,m$。对于四种选择来说，与每月收入增加 1 000 美元分别联系的变化为 0.000，−0.021，0.033 以及−0.012，即岸边、码头、私家船以及租船钓鱼的概率。这表明，岸边钓鱼变动很小，从码头钓鱼及租船钓鱼离开，并向私家船钓鱼方式运动。由于平均月收入是 4 100 美元，所以概率上的变动在合理范围之内。

不过，只有收入对选择钓鱼方式来说不是一个大的辨别因素。由表 15.2 底部可以发现，和 CL 模型相比，MNL 模型具有更小的对数似然以及伪 R^2。从输出不是已知的来看，对于样本中所有不同个体来说，源自 MNL 模型的关于岸边预测概率从 0.095 到 0.115，关于码头的预测概率从 0.036 到 0.234，关于私家船的预测概率从 0.240 到 0.626，而关于租船的预测概率从 0.244 到 0.416。由于 MNL 模型包含截距，这些每个选项的预测概率等于样本平均概率。MNL 模型的这一结果正是稍后式(15.16)给出的结果。

15.2.3 混合 logit

为使模型更为丰富，就要将前面两个模型结合起来。这样做，利用满足

$$\Pr[y_i=j]=\frac{\exp(\beta_P P_{ij}+\beta_C C_{ij}+\alpha_j+\beta_{Ij}I_i)}{\sum_{k=1}^{4}\exp(\beta_P P_{ik}+\beta_C C_{ik}+\alpha_k+\beta_{Ik}I_i)}$$

的所谓混合 logit 模型(参见 15.4.1 节)。不要把该模型与 15.7 节中称为混合模型的那种模型混淆，该模型以条件 logit 模型实施：

$$\Pr[y_i=j]=\frac{\exp(\beta_P P_{ij}+\beta_C C_{ij}+\sum_{l=1}^{4}(\alpha_l d_{ijl}+\beta_{Il} dI_{ijl}))}{\sum_{k=1}^{4}\exp(\beta_P P_{ik}+\beta_C C_{ik}+\sum_{l=1}^{4}(\alpha_l d_{ijl}+\beta_{Il} dI_{ijl}))}$$

其中，d_{ijl} 表示虚拟变量，当 $j=l$ 时，$d_{ijl}=1$，否则为 0；而当 $j\neq l$[1] 时，$dI_{ijl}=d_{ijl}I_i$ 等于收入，否则为 0。在此情况下，我们将 y_i 对 8 个回归元进行回归：P_{ij}、C_{ij}、d_{ij2}、

[1] 原著中这里为等号，怀疑有误，应为不等号。——译者注

d_{ij3}、d_{ij4}、dI_{ij2}、dI_{ij3}以及 dI_{ij4}。由于 $\alpha_1=0$ 且 $\beta_{I1}=0$，所以回归元 d_{ij1} 与 dI_{ij1} 均可以省略。注意到，如果我们估计仅仅以 d_{ijl} 与 dI_{ijl} 作为回归元的这种 CL 模型，那么 CL 估计值等于前面给出的 MNL 估计值。MNL 模型总能够作为 CL 模型得以估计(参见 15.3.4 节)。

尽管混合 logit 模型比 CL 模型更为丰富，但 CL 模型具有下述优点：若额外的选项被添加到选择集合中，则人们能预测选择它的概率，因为 CL 模型的参数并不随选项而变化。

表 15.2 最后一列已报告一些结果。与前面两个模型相比，其系数变动很小，只是捕获率系数变动极大。这种变化归因于包含了特定选项的虚拟变量，而不是因为包含收入。与其他模型相比，混合模型因具有更大的对数似然值或正式统计检验，备受人们青睐。

15.3 一般性结果

本节结果和所有多项式模型有关。本章剩余内容专门研究实际应用中运用的对多项式模型的各种不同设定。

15.3.1 *多项式模型*

存在 m 个模型选项，同时因变量 y 被定义成取 j 值，如果第 j 个选项被采用，$j=1,\cdots,m$。(不过，有些作者考察 $m+1$ 个选项，$j=0,1,\cdots,m$。)将采用第 j 个选项的概率定义成：

$$p_j=\Pr[y=j],\quad j=1,\cdots,m \tag{15.1}$$

对每个观测值 y 引入 m 个二值变量：

$$y_j=\begin{cases}1, & \text{当 } y=j\\ 0, & \text{当 } y\neq j\end{cases} \tag{15.2}$$

因而，y_j 等于 1，若选项 j 是观测结果，而剩下 y_k 等于 0，则对于 y 的每个观测值来说，$y_1,y_2,\cdots,y_m$ 之一将确实是非零的。从而，观测值的多项式密度(**multinomial density**)可方便写成：

$$f(y)=p_1^{y_1}\times\cdots\times p_m^{y_m}=\prod_{j=1}^{m}p_j^{y_j} \tag{15.3}$$

对于回归模型来说，对第 i 个个体及回归元引入下标 i。针对第 i 个个体选择第 j 个选项的概率，建模成：

$$p_{ij}=\Pr[y=j]=F_j(\mathbf{x}_i,\boldsymbol{\beta}),\quad j=1,\cdots,m,\ i=1,\cdots,N \tag{15.4}$$

关于 F_j 的函数形式应该使得概率位于 0 与 1 之间，并对 j 求和为 1。对 F_j 设定各种不同函数就对应于一些特定模型，诸如著名多项式 logit、嵌套 logit、多项式 probit、有序多项式、贯序多项式模型以及多变量模型。这些模型在下面几节加以阐述。

15.3.2 ML 估计

一个观测值的多项式密度已由式(15.3)给出。于是,N 个独立观测值样本的似然函数是 $L_N=\prod_{i=1}^{N}\prod_{j=1}^{m}p_{ij}^{y_{ij}}$,其中,下标 i 表示 N 个个体中的第 i 个,而下标 j 表示 m 个选项中的第 j 个。其对数似然函数(**log-likelihood function**)是:

$$\mathcal{L}=\ln L_N=\sum_{i=1}^{N}\sum_{j=1}^{m}y_{ij}\ln p_{ij} \tag{15.5}$$

其中,$p_{ij}=F_j(\mathbf{x}_i,\boldsymbol{\beta})$ 表示参数 $\boldsymbol{\beta}$ 与回归元的函数,已由式(15.4)定义。更一般地,选项数量会随个体不同而变化,因此,m 选择变成 m_i 选择。

MLE $\hat{\boldsymbol{\beta}}$ 的一阶条件作为

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\beta}}=\sum_{i=1}^{N}\sum_{j=1}^{m}\frac{y_{ij}}{p_{ij}}\frac{\partial p_{ij}}{\partial\boldsymbol{\beta}}=\mathbf{0} \tag{15.6}$$

的解,它通常关于 $\boldsymbol{\beta}$ 是非线性的。y_i 分布一定是多项式的,所以对数据生成过程正确设定意味着,对关于概率 p_{ij} 函数形式 $F_j(\mathbf{x}_i,\boldsymbol{\beta})$ 的正确设定。这就确保了一致性,从而 $\mathrm{E}[y_{ij}]=p_{ij}$,对式(15.6)取数学期望,得到 $\mathrm{E}[\partial\mathcal{L}/\partial\boldsymbol{\beta}]=\sum_{i=1}^{N}\sum_{j=1}^{m}\partial p_{ij}/\partial\boldsymbol{\beta}$,由于$\sum_{j=0}^{m}p_{ij}=1$,因而等于 0。

人们可应用通常渐近理论,从而其方差矩阵为负的信息矩阵的逆。对式(15.6)双和式求关于 $\boldsymbol{\beta}'$ 的微分,并利用 $\mathrm{E}[y_j]=p_{ij}$,得到简化形式:

$$\hat{\boldsymbol{\beta}}\overset{a}{\sim}\mathcal{N}\left[\boldsymbol{\beta}_0,\left(\sum_{i=1}^{N}\sum_{j=1}^{m}\frac{1}{p_{ij}}\frac{\partial p_{ij}}{\partial\boldsymbol{\beta}}\frac{\partial p_{ij}}{\partial\boldsymbol{\beta}'}-\frac{\partial^2 p_{ij}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\bigg|_{\beta_0}\right)^{-1}\right] \tag{15.7}$$

倘若观测值对于不同 i 是独立的,则不要求用更一般方差矩阵的三明治形式,因为数据一定是多项式分布的,而信息矩阵等式将成立。

正如已提及的,各种不同模型对应于 $F_j(\mathbf{x}_i,\boldsymbol{\beta})$ 的不同选择 p_{ij},从而有不同的表达式(15.6)与式(15.7)。关于基于选择样本诸如那些对已观测到结果常常过度抽样的样本的极大似然估计,在 14.5 节与 24.4 节加以阐述。

15.3.3 基于矩的估计

对于简单横截面应用来说,标准估计方法是 MLE。不过,当出现复杂情况,诸如内生性或对不同观测单位 i 具有相关性时,一种更为简便的方法是,使用基于矩的估计量。一旦假定概率得以正确设定,我们考察满足估计方程:

$$\sum_{i=1}^{N}\sum_{j=1}^{m}(y_{ij}-p_{ij})\mathbf{z}_i=\mathbf{0} \tag{15.8}$$

的任何估计量,其中,$\mathbf{z}_i$ 表示与 $\boldsymbol{\beta}$ 的维数相同的向量,$\mathbf{z}_i$ 不依赖于 y_{ij},例如,$\mathbf{z}_i=\partial p_{ij}/\partial\boldsymbol{\beta}$。如果 p_{ij} 的函数形式得到正确设定,那么这个估计量将是一致的,从而 $\mathrm{E}[y_{ij}]=p_{ij}$,而且式(15.8)左边双和式具有 0 期望值。该估计量的有效性将会随 $\mathbf{z}_i$ 选择而变化,而在更一般的情况下,可使用 GMM 估计方法。估计方程(15.8)是多项式 probit 模型的模拟矩方法估计量的基础(参见 15.8.2 节)。

15.3.4 选项变化的回归元

多项式回归模型不仅在式(15.4)关于函数 $F_j(\cdot)$ 的选择方面不同，而且在回归元与参数关于选项方面如何变化上也不同。

在一种极端下，所有回归元都可能是选项变化的(**alternative-varying**)，这意味着对于各种不同选项回归元取不同的值，并设 $\mathbf{x}_i=[\mathbf{x}_{i1}'\ \mathbf{x}_{i2}'\ \cdots\ \mathbf{x}_{im}']$，于是，式(15.4)通常具有

$$F_j(\mathbf{x}_i,\boldsymbol{\beta})=F_j(\mathbf{x}_{i1}'\boldsymbol{\beta},\cdots,\mathbf{x}_{im}'\boldsymbol{\beta})$$

形式，其中，参数 $\boldsymbol{\beta}$ 对不同选项来说为常值。一个例子是后面式(15.10)所定义的条件 logit 模型。

在另一种极端下，所有回归元都可能是选项不变的(**alternative-invariant**)。这意味着，$\mathbf{x}_i$ 并不随选项不同而变化。一个例子是，在交通方式选择模型中的个体社会经济特征。那么，式(15.4)通常具有

$$F_j(\mathbf{x}_i,\boldsymbol{\beta})=F_j(\mathbf{x}_{i1}'\boldsymbol{\beta}_1,\cdots,\mathbf{x}_i'\boldsymbol{\beta}_m)$$

形式，其中，参数 $\boldsymbol{\beta}_j$ 对不同选项会不同，而 $\boldsymbol{\beta}=[\boldsymbol{\beta}_1'\ \boldsymbol{\beta}_2'\ \cdots\ \boldsymbol{\beta}_m']$。参数识别要求正规化，例如 $\boldsymbol{\beta}_1=\mathbf{0}$。一个例子是后面式(15.11)所定义的多项式 logit 模型。

在选项变化回归元与选项不变回归元之间的区别具有重要的实践意义，因为关于多项式模型的标准记号与计算机程序专门地对一种或另一种起作用。当然，在实际应用中，某些回归元可能是选项变化的，而另一些回归元则是选项不变的。在这些情况下，最好是使用为选项变化的回归元编写的程序，因为实施从选项不变回归元到选项变化回归元的格式化是可行的。设 $\mathbf{x}_i$ 表示 $K\times 1$ 维向量。于是，把 $\mathbf{x}_{ij}$ 定义成 $Km\times 1$ 维向量，只有第 j 块为 $\mathbf{x}_i$，其余元素全部为 0，也就是说：

$$\mathbf{x}_{ij}=[\mathbf{0}'\cdots\ \mathbf{0}'\quad \mathbf{x}_i'\quad \mathbf{0}'\cdots\mathbf{0}']$$

并定义 $\boldsymbol{\beta}=[\mathbf{0}'\ \boldsymbol{\beta}_2'\ \cdots\ \boldsymbol{\beta}_m']'$，其中，$\boldsymbol{\beta}_1=\mathbf{0}$ 表示正规化。于是，$\mathbf{x}_i'\boldsymbol{\beta}_j=\mathbf{x}_{ij}'\boldsymbol{\beta}$。本质上，回归元包括与特定选项虚拟变量的交互作用项。一个例子已由 15.2.3 节给出。实施从特定选项回归元到选项不变回归元的格式化同样是可行的，不过需要对每个特定选项回归元施加 $(m-1)$ 个参数等式约束。

15.3.5 显性偏好数据与意向偏好数据

微观经济计量研究所用的多项式数据经常源自个体消费者选择。消费者选择数据，可能是显性偏好数据[1](**revealed preference data**)，即实际决策及结果方面的数据；也可能是意向偏好数据[2](**stated preference data**)，即关于假设方程响应

---

〔1〕 又称为显示性偏好数据。显性偏好理论是由美国著名经济学家保罗・萨缪尔森(P. Samuelson)提出来的，其基本思想是，消费者在一定价格条件下的购买行为暴露或显示他内在的偏好倾向。因此，我们可根据消费者的购买行为来推测消费者的偏好。这是一种不基于"偏好关系(效用函数)—消费者选择"的逻辑思路，而是一个相反的过程，即"消费者选择—偏好关系"。——译者注

〔2〕 又称为叙述性偏好数据或意向调查数据。意向偏好数据是指，其调查内容是尚未发生的事情。意向偏好数据具有如下几个特点：可操作性强、数据误差可调节、意向偏好数据调查中选择方案集合明确等。——译者注

的调查数据。显性偏好数据的一个例子是实际职业的选择。意向偏好数据的例子是关于高效燃料交通工具的市场营销研究，即要求调查对象在诸如燃料消费、范围以及价格特性上不同的各种假设交通工具之间进行选择。

显性偏好数据经常很少提供或没有提供除选择以外的一些选项。例如，我们也许要知道选取产品的个体消费者的价格，而不是可选择产品的价格。用多项式建模的意向偏好数据的引人注目之处是，对于所有可能可选择产品的重要变量诸如价格来说，都具有可利用数据。尤其是，这有助于人们希望预测选择的概率或根据新选项的特性预测该产品的市场份额，如果所有回归元都随选项而变化，那么所有参数关于选项是不变的。

利用意向偏好数据时，存在某种争论，因为响应会随问题措辞而变化。另外，人们可能过分强调，或者少说他们关注支持特殊政策的意愿。例如，一些人愿意过分强调他们支持环境友好政策的意愿。

购物扫描数据(**scanner data**)特别引人注目，因为它们给出展示性选择数据，同时提供各种所有可供选择产品的价格数据。

15.3.6 *模型评价与选择*

对多项式模型中的回归参数直接解释很困难。不过，一种有益方式是，考察回归元变化对结果概率的边际效应(或弹性)。条件 logit 模型与多项式 logit 模型的公式已在 15.4.3 节给出，并在 15.2 节得到了应用。

几种评价模型方法已经由雨宫(Amemiya，1981)与马达拉(Maddala，1983)阐述。利用建立在残差平方类似形式上的 R^2 测量并没有起到很好的作用。将预测概率与实际结果进行比较，得出具有受限制值的特点，因为所估计的含有截距 MNL 模型对估计利用了下述限制：预测概率的平均等于每个选项样本平均概率。考察每个选项的样本内拟合概率的值域是有用的。该值域范围越窄，则越容易辨识模型。对于更详细内容，参见 14.3.7 节的二值结果。

多项式模型通常利用极大似然法进行估计。因而，对于嵌套模型情况来说，运用标准的似然比检验。当模型是非嵌套的时候，运用建立在拟合对数似然上的对模型中参数个数含有自由度调整的赤池信息准则的变形(参见 8.5.1 节)。

归功于麦克法登(McFadden，1973)的有用伪 R^2 测量是：

$$R^2 = 1 - \ln L_{\text{fit}} / \ln L_0 \tag{15.9}$$

其中，$\ln L_{\text{fit}}$表示拟合模型，而 L_0 表示仅有截距的模型，即将每个可供选择的概率估计成为样本平均。对于任何多项式模型来说，对数似然的理论极大值为 0。对于 i 与 j，若当 $y_{ij}=1$，$p_{ij}=1$ 就是此种情况，否则 $p_{ij}=0$。因而，将 R^2 测量重新写成：

$$R^2 = \frac{\ln L_{\text{fit}} - \ln L_0}{\ln L_{\max} - \ln L_0}$$

这能解释成为通过拟合模型所达到的对数似然中最大的潜在增益部分(参见 8.7.1 节)。

15.4 多项式 logit

最简单的多项式模型是多项式 logit 模型，它由卢斯(Luce，1985)提出。广泛运用的此种模型的变形，依据回归元是否随选项不同而变化出现各种形式，本节所阐述的许多问题，与下面几节将更简要讨论的其他模型相联系。

15.4.1 *条件、多项式以及混合 logit 模型*

对于选项变化的回归元(参见 15.3.4 节)来说，运用条件 logit 模型(**conditional logit model**)。CL 模型设定：

$$p_{ij}=\frac{e^{\mathbf{x}_{ij}'\boldsymbol{\beta}}}{\sum_{l=1}^{m}e^{\mathbf{x}_{il}'\boldsymbol{\beta}}},\quad j=1,\cdots,m \tag{15.10}$$

由于 $\exp(\mathbf{x}_{il}'\boldsymbol{\beta})>0$，故这些概率位于 0 与 1 之间，而且对 j 求和为 1。实际上，人们一旦看到公式(15.10)，它看来像确保概率特性良好的一个最简单设定。因为 $\sum_{j=1}^{m}p_{ij}=1$，所以可借助于把 $\mathbf{x}_{ij}$ 定义成回归元与第 1 个选项值的离差，比如说，令 $\mathbf{x}_{i1}=\mathbf{0}$ 来获得等价模型。

不过，当回归元不随选项而变化时，运用多项式 logit 模型(**multinomial logit model**)。MNL 模型设定：

$$p_{ij}=\frac{e^{\mathbf{x}_i'\boldsymbol{\beta}_j}}{\sum_{l=1}^{m}e^{\mathbf{x}_i'\boldsymbol{\beta}_l}},\quad j=1,\cdots,m \tag{15.11}$$

由于 $\sum_{j=1}^{m}p_{ij}=1$，故为了确保模型识别，需要一种约束，而且通常的约束是 $\boldsymbol{\beta}_1=\mathbf{0}$。

上述两个模型能组合成一些作者称为的混合 logit 模型(**mixed logit model**)，它满足：

$$p_{ij}=\frac{e^{\mathbf{x}_{ij}'\boldsymbol{\beta}+\mathbf{w}_i'\boldsymbol{\gamma}_j}}{\sum_{l=1}^{m}e^{\mathbf{x}_{il}'\boldsymbol{\beta}+\mathbf{w}_i'\boldsymbol{\gamma}_l}},\quad j=1,\cdots,m \tag{15.12}$$

其中，$\mathbf{x}_{ij}$ 随选项而变化，而 $\mathbf{w}_i$ 并不随选项而变化。如同 15..2.3 节与 15.3.4 节所讨论的，混合模型与 MNL 模型均能重新表述成 CL 模型。注意，有时候混合 logit 模型术语还用作 15.7 节详述的相当不同的模型。

所有这些模型都能给出一般称谓多项式 logit，但我们遵循标准惯例，对 MNL 模型与 CL 模型加以区别。

对多项式 logit 模型的一种明显推广是：

$$p_{ij}=\frac{V_{ij}}{\sum_{l=1}^{m}V_{il}},\quad j=1,\cdots,m \tag{15.13}$$

其中，$V_{ij}>0$ 可以是回归元 $\mathbf{x}_i$ 与参数 $\boldsymbol{\beta}$ 的相当一般函数。这是所谓的普适 logit 模型[1](**universal logit model**)。尽管这能生成潜在丰富的模型类型，但在经济计量学中却极少使用，因为它并不会由选择理论自然产生。

---

[1] 又称为万能 logit 模型。——译者注

15.4.2 CL与MNL模型的ML估计

我们阐述条件logit模型与多项式logit模型的重要公式。完整推导将在15.12节给出。

对于CL模型来说，其中，p_{ij}已由式(15.10)定义，$\partial p_{ij}/\partial\boldsymbol{\beta}=p_{ij}(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)$，其中，$\bar{\mathbf{x}}_i=\sum_{l=1}^{m}p_{il}\mathbf{x}_{il}$表示回归元的概率加权平均(参见15.12.1节)CL一阶条件，即由式(15.6)给出一般p_{ij}，可立刻简化成：

$$\sum_{i=1}^{N}\sum_{j=1}^{m}y_{ij}(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)=\mathbf{0} \tag{15.14}$$

对$\boldsymbol{\beta}'$求导数，利用$\mathrm{E}[y_{ij}]=p_{ij}$，并经过某种进一步代数运算，得到：

$$\hat{\boldsymbol{\beta}}_{\mathrm{CL}}\overset{a}{\sim}\mathcal{N}\Big[\boldsymbol{\beta},\ \Big(\sum_{i=1}^{N}\sum_{j=1}^{m}p_{ij}(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)'\Big)^{-1}\Big] \tag{15.15}$$

对于MNL模型来说，p_{ij}已由式(15.11)定义，而且15.12.2节将证明，$\partial p_{ij}/\partial\boldsymbol{\beta}_k=p_{ij}(\delta_{ijk}-p_{ik})\mathbf{x}_i$，其中，$\delta_{ijk}$表示指示变量，当$j=k$时，$\delta_{ijk}$等于1，而当$j\neq k$时，$\delta_{ijk}$等于0，并且得到的MNL一阶条件在经过某些代数运算之后，简化成：

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\beta}_k}=\sum_{i=1}^{N}(y_{ik}-p_{ik})\mathbf{x}_i=0,\quad k=1,\cdots,m \tag{15.16}$$

正常情况下，$\hat{\boldsymbol{\beta}}_{\mathrm{MNL}}\overset{a}{\sim}\mathcal{N}[\boldsymbol{\beta},(\mathrm{E}[\partial^2\mathcal{L}/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'])^{-1}]$，经过进一步代数运算可以证明，信息矩阵的第$jk$个块为：

$$\mathrm{E}\Big[\frac{\partial^2\mathcal{L}}{\partial\boldsymbol{\beta}_j\partial\boldsymbol{\beta}_k'}\Big]=\sum_{i=1}^{N}p_{ij}(\delta_{ijk}-p_{ik})\mathbf{x}_i\mathbf{x}_i',\quad j=1,\cdots,m,\ k=1,\cdots,m \tag{15.17}$$

15.4.3 回归参数解释

在任何非线性模型中，对参数进行解释都需要小心慎重。对多项式模型而言，尤其如此，例如，在系数符号与系数概率之间不一定存在一一对应。这里，我们阐述在15.2节的应用中使用的结果。

边际效应与弹性

我们关注给定个体时回归元变化对选择概率的边际效应(**marginal effects**)。于是，弹性(**elasticities**)能通过利用当前回归元乘以边际效应，并用概率去除而计算出。典型地讲，为了给出平均边际效应或平均弹性，这是关于个体的平均。

对于CL模型来说，考察关于第k个选项的回归元变动1个单位时对第j个概率的效应。例如，如果乘公共汽车旅行的时间增加1分钟，而通过其他方式旅行的时间不改变，那么选择各种运输方式的效应是什么呢？由15.12.1节知：

$$\frac{\partial p_{ij}}{\partial\mathbf{x}_{ik}}=p_{ij}(\delta_{ijk}-p_{ik})\boldsymbol{\beta} \tag{15.18}$$

其中，δ_{ijk}已在式(15.15)后面定义。由此可得，如果回归元系数是正的，那么关于

第 k 个选项的相对应值的回归元的成分增大，会增加第 k 个选项的概率，同时减少其他选项的概率。

然而，对于 MNL 模型来说，考察对所有选项都取相同值的回归元变动一个单位时第 j 个概率的效应。例如，年龄增大 1 年对选取工作的概率效应是什么？由 15.12.2 节知：

$$\frac{\partial p_{ij}}{\partial \mathbf{x}_i}=p_{ij}(\boldsymbol{\beta}_j-\bar{\boldsymbol{\beta}}_i) \tag{15.19}$$

其中，$\bar{\boldsymbol{\beta}}_i=\sum_l p_{il}\boldsymbol{\beta}_l$ 表示 $\boldsymbol{\beta}_l$ 的概率加权平均值。由此可得，响应符号不一定是由 $\boldsymbol{\beta}_j$ 的符号给出，除非 $\boldsymbol{\beta}_j>\boldsymbol{\beta}_k$，对于所有 $k\neq j$，同时不一定要检验特定系数是否为 0。如同其他非线性模型一样，我们可计算平均响应 $N^{-1}\sum_i\partial p_{ij}/\partial\mathbf{x}_i=N^{-1}\sum_i p_{ij}(\boldsymbol{\beta}_j-\bar{\boldsymbol{\beta}}_i)$，或者使用非微分方法，并比较当回归元变动时平均预测概率的变化。

基准类的比较

CL 模型与 MNL 模型中的系数同样能依据(14.3.4 节详细阐述的)相对风险给出更直接的如同 logit 的解释。这是因为该模型可重新表述成二值 logit 模型。

对于 MNL 模型来说，比较是针对基准类，这是选项正规化拥有等于 0 的系数。为了认识这一类，注意到，如果选项 j 或选项 k 可观测，多项式 logit 概率(15.11)蕴含，可观测选项 j 的条件概率是：

$$\begin{aligned}\Pr[y=j\,|\,y=j\text{ 或 }k]&=\frac{p_j}{p_j+p_k}\\&=\frac{e^{\mathbf{x}'\boldsymbol{\beta}_j}}{e^{\mathbf{x}'\boldsymbol{\beta}_j}+e^{\mathbf{x}'\boldsymbol{\beta}_k}}\\&=\frac{e^{\mathbf{x}'(\boldsymbol{\beta}_j-\boldsymbol{\beta}_k)}}{1+e^{\mathbf{x}'(\boldsymbol{\beta}_j-\boldsymbol{\beta}_k)}}\end{aligned} \tag{15.20}$$

这是具有系数$(\boldsymbol{\beta}_j-\boldsymbol{\beta}_k)$的 logit 模型。经过某种简化可得第二个等式。假定对选项 1 进行正规化，所以 $\boldsymbol{\beta}_1=\mathbf{0}$。于是有：

$$\Pr[y_i=j\,|\,y_i=j\text{ 或 }1]=\frac{e^{\mathbf{x}_i'\boldsymbol{\beta}_j}}{1+e^{\mathbf{x}_i'\boldsymbol{\beta}_j}}$$

以同样方式对 $\boldsymbol{\beta}_j$ 解释成在选项 j 与 1 之间二值选择的 logit 模型系数。类似于二值 logit 模型，选择选项 j 而不是选项 1 的相对风险是：

$$\frac{\Pr[y_i=j]}{\Pr[y_i=1]}=e^{\mathbf{x}_i'\boldsymbol{\beta}_j}$$

因而 $e^{\beta_{jr}}$ 给出，当 x_{ir} 变化一个单位时，这种相对风险中的比例变化。这种解释将会依据哪一个选项被正规化成拥有零系数而变化，并且人们需要拥有一种自然的基准类(**base category**)，这种解释确实是有用的。例如，倘若关注内容在于各种可选择的旅行汽车往返方式，就对汽车选项的系数正规化成 0。

同理，类似方式用于满足：

$$\Pr[y_i=j\,|\,y_i=j\text{ 或 }k]=\frac{e^{(\mathbf{x}_{ij}-\mathbf{x}_{ik})'\boldsymbol{\beta}}}{1+e^{(\mathbf{x}_{ij}-\mathbf{x}_{ik})'\boldsymbol{\beta}}} \tag{15.21}$$

的 CL 模型，而现在正规化是针对基准类回归元值进行的。

15.4.4 *无关选项的独立性*

CL 模型与 MNL 模型的局限性是，在 m 个选项之间进行辨别就被简化成一系列两两比较，这种两两比较除了所考虑的两两对比之外没有受到选项特征的影响。由式(15.20)与式(15.21)知，这是很明显的，可以证明，MNL 模型简化成任何选择对之间的二值选择 logit 模型。该条件概率并不依赖于其他选项。

举一个极端例子，给定乘小车或红色公共汽车往返两地，MNL 模型或 CL 模型中往返两地的条件概率被假定成与是否乘蓝色公共汽车往返的选项是独立的。不过，实际上我们希望引进蓝色公共汽车，除颜色之外，蓝色公共汽车在每个方面都与红色公共汽车一样，很少对小车使用产生影响，同时将红色公共汽车的使用减半，导致了给定乘小车或红色公共汽车往返时对小车使用的条件概率增大。

MNL 的这一弱点，在文献上统称为红色公共汽车—蓝色公共汽车问题，或更正式地，称为无关选项的独立性[1](**independence of irrelevant alternatives**)。利用豪斯曼检验可对它进行检验[参见豪斯曼和麦克法登(Hausman and McFadden, 1984)]。例如，我们能计算出小车、红色公共汽车以及蓝色公共汽车的三种选择模型中红色公共汽车的系数估计值，这里再次以小车为基准类，与系数估计值加以比较。大多数经济计量学文献都关注于没有这种弱点的可选择无序模型。这些模型将在 15.6 节至 15.8 节阐述。

15.5 可加随机效用模型

比多项式 logit 与条件 logit 模型更为一般的无序多项式模型，通过利用可加随机效用模型的一般框架来获得，本节阐述可加随机效用模型。下面几节阐述重要例子。

15.5.1 *ARUM*

14.4.2 节已经引入二值结果的可加随机效用模型(**additive random utility model**)。在一般的 m 个选择多项式模型中，第 j 个选择的效用被设定成：

$$U_j=V_j+\varepsilon_j,\quad j=1,2,\cdots,m \tag{15.22}$$

其中，V_j 表示效用的确定性成分，而 ε_j 表示效用的随机成分。对于第 i 个个体来说，通常是 $V_{ij}=\mathbf{x}_{ij}'\boldsymbol{\beta}$ 或 $V_{ij}=\mathbf{x}_i'\boldsymbol{\beta}_j$，通过另外的结构分析，可设定消费者需求理论使用的直接或间接效用函数。为了记号简单起见，下面不用个体下标 i。

被选择的选项是具有最大效用的，所以：

$$\begin{aligned}\Pr[y=j]&=\Pr[U_j\geqslant U_k,\ 所有\ k\neq j]\\&=\Pr[U_k-U_j\leqslant 0,\ 所有\ k\neq j]\\&=\Pr[\varepsilon_k-\varepsilon_j\leqslant V_j-V_k,\ 所有\ k\neq j]\\&=\Pr[\tilde{\varepsilon}_{kj}\leqslant \tilde{V}_{kj},\ 所有\ k\neq j]\end{aligned} \tag{15.23}$$

[1] 又称为无关选择的独立性。——译者注

其中，"～"与另一个下标 j 表示针对参照选项 j 的微分。

各种多项式模型可通过误差项联合分布的不同假设生成。这些模型在统计上是有效的，因为概率之和为 1。另外，模型与决策的标准经济理论相一致。

例如，考察三种选择模型中的 Pr[y=1]的表达式。利用式(15.23)中的最后一个等式，并定义 $\tilde{\varepsilon}_{31}=\varepsilon_3-\varepsilon_1$，$\tilde{\varepsilon}_{21}=\varepsilon_2-\varepsilon_1$，则有：

$$\begin{aligned}\Pr[y=1] &= \Pr[\tilde{\varepsilon}_{21}\leqslant-\tilde{V}_{21},\ \tilde{\varepsilon}_{31}\leqslant-\tilde{V}_{31}] \\ &= \int_{-\infty}^{-\tilde{V}_{31}}\int_{-\infty}^{-\tilde{V}_{21}} f(\tilde{\varepsilon}_{21},\tilde{\varepsilon}_{31})d\tilde{\varepsilon}_{21}d\tilde{\varepsilon}_{31}\end{aligned} \tag{15.24}$$

这是一个二变量积分，通常没有解析解。更一般地，m 种选择模型会涉及($m-1$)变量积分，该积分可能产生 Pr[$y=j$]的闭形式解，也可能没有 Pr[$y=j$]的闭形式解。

通常，所有误差对于不同选择来说可能是相关的。不过，需要某些协方差约束，因为模型是可识别的，只是至多相差($m-1$)个误差差分对[参见式(15.23)的最后一个等式]，同时由于 U_j 仅仅至多相差一个标度是确定的，所以需要设定一个方差。

15.5.2 *各种各样无序多项式模型*

各种无序多项式模型起因于对 $\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_m$ 联合分布的不同假设。若误差假设导致选择概率的闭形式解，分析就相当简单。不过，在许多应用中，这些假设被认为约束性太强。

即使选择概率不存在闭形式解，第 12 章已归纳的密集计算法使得该估计变得容易。15.7.2 节与 15.8.2 节将阐述这些方法的多项式例子。

第 1 类型极值误差

首先假定，误差 ε_j 均是 iid 的且为第 1 类型极值误差，其密度为：

$$f(\varepsilon_j)=e^{-\varepsilon_j}\exp(-e^{-\varepsilon_j}),\quad j=1,2,\cdots,m \tag{15.25}$$

该密度性质已由 14.4.2 节给出，那里已经证明，在二值结果情况下，这就产生 logit 模型。

对于利用含有第 1 类型极值误差的 ARUM 进行建模的多项式结果来说，可以证明，式(15.23)导致：

$$\Pr[y=j]=\frac{e^{V_j}}{e^{V_1}+e^{V_2}+\cdots+e^{V_m}} \tag{15.26}$$

当 $V_j=\mathbf{x}_j'\boldsymbol{\beta}$ 时，这是一个 CL 模型，而当 $V_j=\mathbf{x}'\boldsymbol{\beta}_j$ 时，这是一个 MNL 模型。该结果可通过积分且类似于二值情况的简化来获得(参见 14.8 节)，或者作为 15.6 节推导的嵌套 logit 结果的一种特殊情况。因此，条件 logit 与多项式 logit 模型可从 ARUM 获得。

误差 ε_j 对于不同选项 j 来说是独立的这个假设，表现得约束性太强，因为若两个选项类似，可能就会违背它。例如，假定选项 1 与选项 2 是类似的。ε_1 的很小值(也就是说，大的且负的)会导致对选项 1 效用的过度预测。随后，我们同样会过度

预言选项 2 的效用，所以 ε_2 也取很小值。由于 ε_1 与 ε_2 的很小值会趋于一致，同时对于很大值来说，类似地，误差必是相关的。这是以另一种方式看待“红色公共汽车—蓝色公共汽车”问题，而且它是 logit 无关选项的独立性假设失败的证明。

广义极值模型与嵌套 logit 模型(参见 15.6 节)都放松了极值误差对不同选择是独立的假设。误差以不同组具有独立性而得以分组，却允许组内相关。于是，对选择概率来说，可利用闭形式解。尽管这些模型比 MNL 模型组内无关的特殊情况更为丰富，但在许多应用中，对误差分组显然有点任意性。

随机参数 logit 模型(参见 15.7 节)将可加随机性引入导致效用对不同选项相关的 MNL 模型之中。这是广义随机效用模型的例子(参见 15.7.3 节)。

正态分布误差

如果假定误差 $\varepsilon_1,\cdots,\varepsilon_m$ 服从联合正态分布，就得出多项式 probit 模型(参见 15.8 节)。与第 I 类型极值的这种误差假设相比，是更为自然的起点。它允许出现非常丰富的相关结构，只是以需要使用数值方法或模拟方法作为代价，而这两种方法都适应 $(m-1)$ 变量正态分布。

15.5.3 *随机效用模型与一致性*

阐述选择概率位于 0 与 1 之间的解析表达式，同时对选项求和为 1 总是可能的。一种相当一般的例子是普适 logit 模型(15.13)。经济计量学文献极为重视多项式模型，多项式模型与对随机效用函数求最大值相一致。这类似于对需求函数的限制分析，此种需求函数与消费者选择理论相一致。

设 $V=(V_1,\cdots,V_m)$。由伯尔施—祖潘(Borsch-Supan, 1987, 第 19 页)知，一组选择概率 $p_j(V)$ 与对 ARUM 求最大值并不矛盾，$j=1,\cdots,m$，如果：

1. 对于所有 $\alpha\in R$，$p_j(V)\geqslant 0$，$\sum_{j=1}^{m}p_j(V)=1$，$p_j(V)=p_j(V+\alpha)$；

2. $\partial p_j(V)/\partial V_k=\partial p_k(V)/\partial V_j$；

3. $\partial^{(m-1)}p_j(V)/\partial V_1\cdots[\partial V_i]\cdots\partial V_m\geqslant 0$，其中，方括号表示被省略的项。

这些条件归功于威廉姆(William, 1977)、戴利和扎卡里(Daly and Zachary, 1979)以及麦克法登(McFadden, 1981)。该条件依次确保：(1) 特性良好的概率与变换不变性；(2) p_j 的可积性类似于斯卢茨基(Slutsky)条件；(3) 对应于 ARUM 中误差的分布函数具有正常(非负的)密度函数。

15.5.4 *福利分析*

利用多项式模型的一个主要优点是，可用随机效用模型进行福利分析。于是，人们能对选择的一个或多个决定因素的变动效应赋予美元价值，诸如在交通方式选择方面的旅行价格或者时间成本。

标准的福利分析(**welfare analysis**)运用补偿变化或等价变化。式(15.22)的确定性效用成分被设定成间接效用函数：

$$V_j=V(I-p_j,\ \mathbf{x}_j) \tag{15.27}$$

其中，I 表示收入，p_j 表示第 j 个选项的价格，而 $\mathbf{x}_j$ 表示与第 j 个选项联系的特征。

为了记号简单，不用未知回归参数 $\boldsymbol{\beta}$。选项 j 的效用是：

$$U_j=U(I-p_j, \mathbf{x}_j, \varepsilon_j)=V(I-p_j, \mathbf{x}_j)+\varepsilon_j \tag{15.28}$$

假定我们变动特征从 $\mathbf{x}_j'$ 到 $\mathbf{x}_j''$。然后，补偿变化(**compensating variation**)CV 是为使效用保持在其最初水平上而需要收入的变动，因此，具有收入 I 与特征 $\mathbf{x}_j'$ 的可达最大效用水平必须等于具有收入($I-CV$)与特征 $\mathbf{x}_j''$ 的可达最大效用水平。因而，补偿变化 CV 可以用隐性方式定义成下式的结果：

$$\max_{j=1,\cdots,m} U(I-p_j, \mathbf{x}_j', \varepsilon_j)=\max_{j=1,\cdots,m} U(I-CV-p_j, \mathbf{x}_j'', \varepsilon_j) \tag{15.29}$$

举一个例子，考察两个选择的模型，其中，$U_j=I+x_j+\varepsilon_j$，$j=1,2$，而且纯量 x_j 变动从 x_j' 到 x_j''。于是，存在四种可能性。如果对选项 1 被选取前后进行对比，那么 $CV=(x_1''-x_1')$，从而 $U_1''=I-CV+x_1''+\varepsilon_1=I+x_1'+\varepsilon_1=U_1'$。类似地，如果对选项 2 被选取前后进行对比，那么 $CV=(x_2''-x_2')$。如果发生从选项 1 到 2 的变动，那么 $U_2''=U_1'$ 蕴含 $I-CV+x_2''+\varepsilon_2=I+x_1'+\varepsilon_1$，这蕴含 $CV=x_2''-x_1'+\varepsilon_2-\varepsilon_1$。类似地，如果发生从选项 2 到选项 1 的变动，那么 $CV=x_1''-x_2'+\varepsilon_1-\varepsilon_2$。更一般地，对于 m 个选择来说，如果 x 变化导致从选项 j 到选项 k 的变动，那么在此样本例子中，补偿变化是 $CV_{jk}=V_k''-V_j'+\varepsilon_k-\varepsilon_j$。

补偿变化依赖于可观测值(I, p_j 以及 $\mathbf{x}_j$)、可加以估计的参数以及不可观测的误差 ε_j。不可观测的因素可通过计算期望补偿变化 E[CV]加以剔除，这涉及对 ε_j 进行积分。由前面例子知，应该很明显，这个积分相当难计算。达格斯文科和卡尔斯特罗姆(Dagsvik and Karlström, 2004)曾提供相当一般的结果，15.6.5 节将进一步讨论。

对于某些模型来说，E[CV]不存在解析解。人们转而需要对式(15.29)所定义的关于 CV 的 ε_j 函数进行数值积分。由 12.3.2 节知，此积分能以下述方式进行模拟：

1. 对于源自 $\varepsilon=(\varepsilon_1,\cdots,\varepsilon_m)$分布的 s 个采样 ε^s 进行迭代。
2. 由 $\max_{j=1,\cdots,m} U(I-p_j, \mathbf{x}_j', \varepsilon_j^s)=\max_{j=1,\cdots,m} U(I-CV^s-p_j, \mathbf{x}_j'', \varepsilon_j^s)$，计算出 CV^s。
3. 重复第一步与第二步 S 次。
4. 利用 $S^{-1}\sum_{s=1}^{S} CV^s$估计 E[$CV$]。

对于样本中的每一个个体，该方法都会得到 E[CV]。一旦进行平均，可能利用那个权数，则得到总体估计。15.6.5 节将讨论 GEV 模型的一个应用。

15.6 嵌套 logit

嵌套 logit 是多项式模型在解析形式上最容易处理的推广。当存在明显嵌套结构时，嵌套 logit 是一种理想的模型，但并不是所有的多项式选择应用都具有明确的嵌套结构。

15.6.1 *广义极值模型*

麦克法登(McFadden, 1978)曾提出建立在下述假设之上相当一般的模型类

别，该假设为误差的联合分布是具有联合分布函数：

$$F(\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_m)=\exp[-G(e^{-\varepsilon_1},e^{-\varepsilon_2},\cdots,e^{-\varepsilon_m})] \tag{15.30}$$

的广义极值(**generalized extreme value**，记为 **GEV**)分布，其中，$G(Y_1,Y_2,\cdots,Y_m)$ 函数被设定成满足一系列假设：包括非负性，自由度为 1 的齐性，偶数阶的混合偏导数为连续的且非正的，而奇数阶混合偏导数为非负的，同时 $\lim_{Y_j\to\infty}G(Y_1,Y_2,\cdots,Y_m)=\infty$。这些假设确保了，联合分布与所得到的边缘分布都是良好定义的且概率之和为 1。

倘若误差服从 GEV 分布，就能获得随机效用模型(15.22)中的概率显性解，其满足：

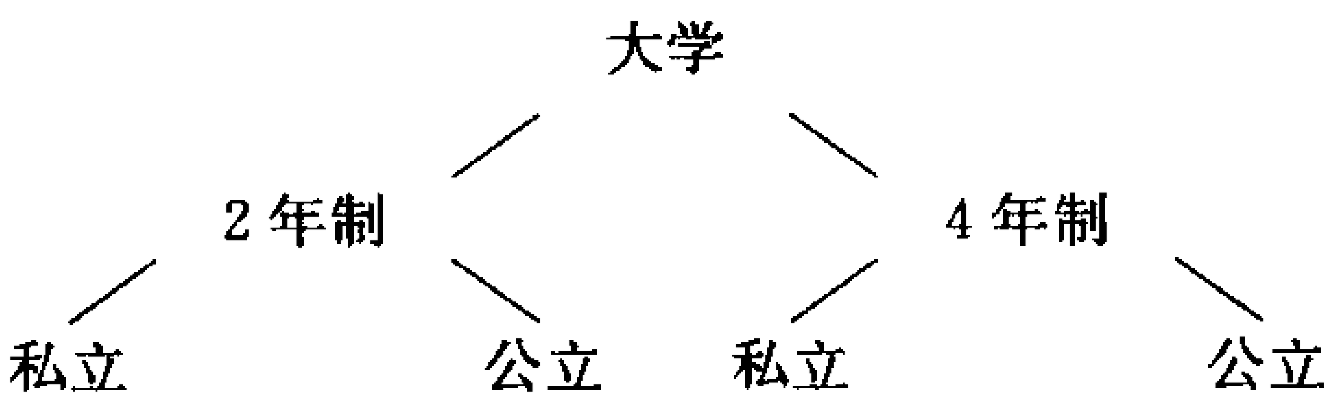
$$p_j=\Pr[y=j]=e^{V_j}\frac{G_j(e^{-V_1},e^{-V_2},\cdots,e^{-V_m})}{G(e^{-V_1},e^{-V_2},\cdots,e^{-V_m})} \tag{15.31}$$

其中，$G_j(Y_1,Y_2,\cdots,Y_m)=\partial G(Y_1,Y_2,\cdots,Y_m)/\partial Y_j$[参见麦克法登(McFadden，1978，第 81 页)]。

通过对 $G(Y_1,Y_2,\cdots,Y_m)$ 的不同选取，就能获得广泛的模型。当 $G(Y_1,Y_2,\cdots,Y_m)=\sum_{k=1}^{m}Y_k$ 时，可获得 MNL 模型，因此，MNL 模型是一种 GEV 模型。另一种广泛运用的 GEV 模型是嵌套 logit 模型。

15.6.2 *嵌套 logit 模型*

嵌套 logit 模型将决策分成一些组。一个简单例子是，考察对大学的选择，其中人们首先决策是否上两年制或四年制大学，然后对这两种路径中的每一组决定是上公立的还是私立的大学。对这种情形作图说明如下：

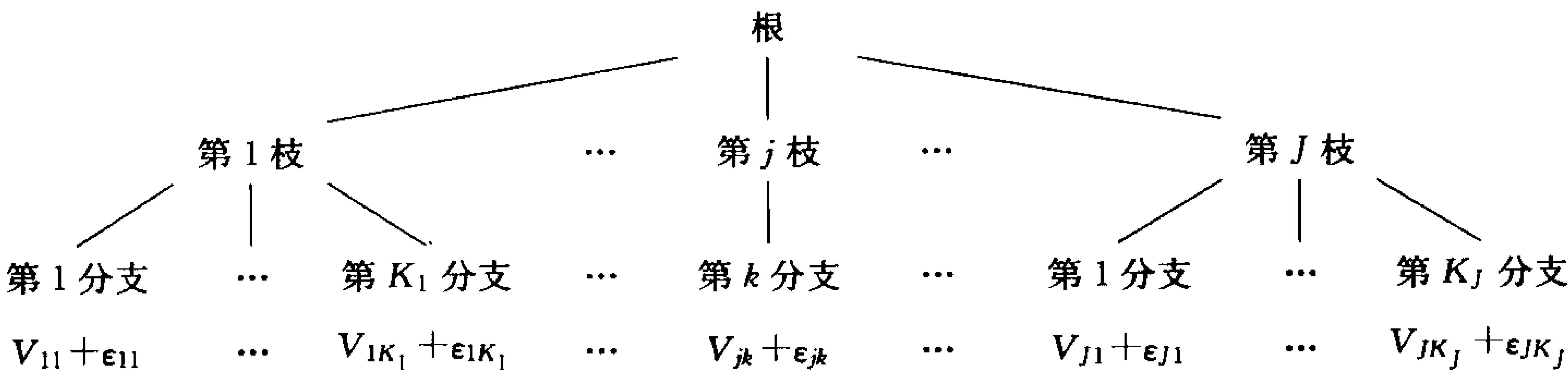


在两年制组与四年制组的每个组内，允许随机效用模型的误差对每个选项是相关的，但对两个组来说，误差是不相关的。

更一般地讲，我们假定在顶部水平存在 J 个要选择的枝。第 j 枝具有 K_j 个分支，它们记为 $j1,\cdots,jk,\cdots,jK_j$。于是，对于 J 个枝中第 j 枝且 K_j 个分支中第 k 分支选项来说，其效用是：

$$U_{jk}=V_{jk}+\varepsilon_{jk},\quad k=1,2,\cdots,K_j,\ j=1,2,\cdots,J \tag{15.32}$$

其中对于 m 种选择模型而言，$K_1+\cdots+K_J=m$。这可阐明如下：



可能存在另一些水平，其中第三个水平是细枝等。为了记号简单起见，我们阐述两水平模型的结果。

对于含有这种嵌套的任何模型来说，作为第 j 枝且第 k 分支的联合概率能被分解因子 p_{jk}——选择第 j 枝的概率——乘以以第 j 枝为条件选择第 k 分支的概率。因而，有：

$$p_{jk}=p_j\times p_{k|j}$$

当误差项 ε_{jk} 具有 GEV 联合累积分布函数：

$$F(\varepsilon)=\exp[-G(e^{-\varepsilon_{11}},\cdots,e^{-\varepsilon_{1K_1}};\cdots;e^{-\varepsilon_{J1}},\cdots,e^{-\varepsilon_{JK_J}})] \tag{15.33}$$

对于函数 $G(\cdot)$ 的下述特殊设定

$$G(\mathbf{Y})=G(Y_{11},\cdots,Y_{1K_1},\cdots,Y_{J1},\cdots,Y_{JK_J})=\sum_{j=1}^{J}(\sum_{k=1}^{K_j}Y_{jk}^{1/\rho_j})^{\rho_j} \tag{15.34}$$

来说，就产生麦克法登(McFadden, 1978)嵌套 logit 模型。参数 ρ_j 表示 ε_{jk} 与 ε_{jl} 之间的相关函数，但不精确地等于相关参数。实际上，可以证明，ρ_j 等于 $\sqrt{1-\mathrm{Cor}[\varepsilon_{jk},\varepsilon_{jl}]}$，因此，$\rho_j$ 反过来与相关性有关，而且我们希望 $0\leqslant\rho_j\leqslant1$。选取 $\rho_j=1$ 对应于 ε_{jk} 与 ε_{jl} 的独立性，从而导致 MNL 模型。我们称参数 ρ_j 为标度参数[1]（**scale parameter**），因为它们对下述所要考察模型的回归参数进行标度。

记号会因作者不同而出现相当大的变化。麦克法登(McFadden, 1978)与马达拉(Maddala, 1983)却以 $\sigma_j=1-\rho_j$ 来定义这个 cdf，称为非相似参数（**dissimilarity parameter**）。另外一些作者则使用 $\mu_j=1/\rho_j$。许多作者对第 n 个体的选项 ij 进行建模，然而我们对选项 jk 进行建模，同时把 i 用于第 i 个个体。

当选项 jk 被选上时，y_{jk} 结果指示变量等于 1，否则为 0。然后，由式(15.32)知，$p_{jk}=\Pr[y_{jk}=1]=\Pr[U_{jk}\geqslant U_{lm}$，对于所有 $l,m]$。作为 V_{jk} 与 ρ_j 函数的概率 p_{jk} 的闭形式解将在 15.12.3 节加以推导。于是，对特殊的确定效用函数：

$$V_{jk}=\mathbf{z}_j'\boldsymbol{\alpha}+\mathbf{x}_{jk}'\boldsymbol{\beta}_j,\ k=1,\cdots,K_j,\ j=1,\cdots,J \tag{15.35}$$

计算这些值，其中，$\mathbf{z}_j$ 仅仅随着枝而变化，而 $\mathbf{x}_{jk}$ 则既随枝又随分支不同而变化。参数 $\boldsymbol{\alpha}$ 与 $\boldsymbol{\beta}_j$ 称为回归参数（**regression parameters**）。

GEV 模型(15.32)～(15.35)会产生嵌套 logit 模型（**nested logit model**）：

$$p_{jk}=p_j\times p_{k|j}=\frac{\exp(\mathbf{z}_j'\boldsymbol{\alpha}+\rho_jI_j)}{\sum_{m=1}^{J}\exp(\mathbf{z}_m'\boldsymbol{\alpha}+\rho_mI_m)}\times\frac{\exp(\mathbf{x}_{jk}'\boldsymbol{\beta}_j/\rho_j)}{\sum_{l=1}^{K_j}\exp(\mathbf{x}_{jl}'\boldsymbol{\beta}_j/\rho_j)} \tag{15.36}$$

参见 15.12.3 节，其中：

$$I_j=\ln(\sum_{l=1}^{K_j}\exp(\mathbf{x}_{jl}'\boldsymbol{\beta}_j/\rho_j)) \tag{15.37}$$

称为相容值（**inclusive value**）或者对数和（**log-sun**）。嵌套 logit 模型的引人注目之处是，概率 p_i 与 $p_{j|i}$ 本质上都具有条件 logit 形式。

---

〔1〕 又称为尺度参数。——译者注

前面结果是关于对不同选项都是变化的回归元的。经过一些代数运算，可适应于选项不变回归元 $V_{jk}=\mathbf{z}'\boldsymbol{\alpha}_j+\mathbf{x}'\boldsymbol{\beta}_{jk}$，对 $\boldsymbol{\beta}_{jk}$ 进行正规化为 1。所需做的全部内容，在代数形式上就是划分 $V_{jk}=A_j+B_{jk}$，其中，A_j 涉及枝，而 B_{jk} 既涉及枝又涉及分支。

15.6.3 *嵌套 logit 的估计*

对于第 i 个观测值来说，我们可观测到 $K_1+\cdots+K_J$ 个结果，当选项 jk 被选取时，$y_{ijk}=1$，否则 $y_{ijk}=0$。于是，$p_{ijk}=p_{ik|j}\times p_{ij}$，而且观测值 $\mathbf{y}_i=(y_{i11},\cdots,y_{iJK_J})$ 的密度能以简洁方式表述成：

$$f(\mathbf{y}_i)=\prod_{j=1}^{J}\prod_{k=1}^{K_j}[p_{ik|j}\times p_{ij}]^{y_{ijk}}=\prod_{j=1}^{J}\left(p_{ij}^{y_{ij}}\prod_{k=1}^{K_j}p_{ik|j}{}^{y_{ijk}}\right)$$

其中，当枝 j 被选取时，$y_{ij}=\sum_{l=1}^{K_j}y_{ijl}$ 等于 1，否则 y_{ij} 为 0。

关于样本密度是 $\prod_{i=1}^{N}f(\mathbf{y}_i)$。FIML 估计量（**FIML estimator**）对参数 $\boldsymbol{\alpha}$、$\boldsymbol{\beta}_j$ 以及 ρ_j 求

$$\ln L=\sum_{i=1}^{N}\sum_{j=1}^{J}y_{ij}\ln p_{ij}+\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K_j}y_{ijk}\ln p_{ik|j}\tag{15.38}$$

的最大值。

一种可供选择的方式，较少有效的估计是序贯估计量（**sequential estimator**）或者 LIML 估计量，而 LIML 估计量利用了把 p_{jk} 分成 $p_{k|j}$ 与 p_j 的乘积。第一阶段估计是根据式（15.38）右边第二项来进行的，由式（15.36）知，这是含有估计参数 $\boldsymbol{\beta}_j/\rho_j$ 的条件 logit。第二阶段估计是根据右边第一项来进行的，由式（15.36）知，这是含有添加回归元 $\hat{I}_{ij}$ 的条件 logit 模型，式（15.37）中相容值的估计值可利用第一阶段参数估计值计算出来。从第二阶段可直接获得 $\hat{\boldsymbol{\alpha}}$ 与 $\hat{\rho}_j$，而 $\hat{\boldsymbol{\beta}}_j$ 等于 $\hat{\rho}_j$ 乘以第一阶段估计值 $\widehat{\boldsymbol{\beta}_j/\rho_j}$。

与 FIML 估计量相比，这种序贯估计量的有效性稍差一些，而且在第二阶段，通常 CL 标准误差低估了序贯估计量的真实标准误差，因为它们并没有考虑计算相容值的估计误差。麦克法登（McFadden，1981）曾给出校正标准误差的公式，或者使用自助法。每当条件 logit 模型估计遇到挑战时，最初就要提出一种其他可供选择的序贯估计量。现在，对似然函数编程来说相对简单，所以最好是使用 FIML。序贯估计潜在地有助于提供初值，因为 FIML 对数似然不是全局凹的。

举一个例子，我们把嵌套 logit 模型应用到 15.2 节的数据上。嵌套结构在较高水平上是岸上钓鱼或船上钓鱼，其较低水平是岸边或码头（岸上钓鱼）以及私家船或租船（船上钓鱼）。式（15.36）中，在较低水平变化的回归元是价格（P）与捕获率（C）。在较高水平变化的回归元 $\mathbf{z}_j$ 对于岸上或船上来说是一个指示变量 d，当在岸上钓鱼时有 $d=1$，而 $d\times I$ 表示收入与岸上钓鱼指示变量的交互作用。通过条件 logit（对应于 $\rho_1=\rho_2=1$）加以估计，得到带有 $\ln L=-1\,252$ 的拟合模型，如同预期的，比类似的对数似然要小一些，只是不及由表 15.2 的最后一列给出的约束模型。对应的嵌套 logit 模型的 FIML 估计，其中，ρ_1 与 ρ_2 现在可自由变化，会产生

更大一些的对数似然模型，并利用 $\chi^2(2)$ 似然比检验统计量对更有约束性的条件 logit 模型拒绝。

15.6.4 讨论

嵌套 logit 模型的主要局限性是，不是所有的选择问题都具有明显嵌套结构。人们还能利用似然比检验或赤池信息准则（尽可能适当）选择最优嵌套方案。不过，所得到的方案并不总是与先验预期相吻合。

另一个实际问题是，含有源自 ARUM 选择的嵌套 logit 模型的一致性，需要 15.5.2 节的三个条件都得以满足。这些条件中的第 i 个会全局性地得以满足，当 $0 \leqslant \rho_j \leqslant 1$ 时，对具有多于两个水平的嵌套来说，它会另外要求在嵌套结构较高水平的 ρ 不大于嵌套的较低水平的 ρ。在实际应用中，获得位于单位区间之外的 ρ_j 估计值是可能的。由于选择概率是正常的，所以人们还能够运用此模型，只是模型不再来自 ARUM。伯尔施—祖潘以及一些其他人曾经考察了，在嵌套 logit 模型可以与 ARUM 一致的条件下的局部识别条件，即使 ρ_j 位于单位区间之外。为了把 ρ_j 限制到单位区间上，且统计出对数似然的减少，若有的话，这样做要小心谨慎，对 ρ_j 进行格点搜索是有用的。

由式(15.36)与式(15.37)定义的嵌套 logit 模型是由麦克法登(McFadden，1978)提出的，他把它推导成 GEV 模型。嵌套 logit 模型的较早变形(**earlier variant**)类似于式(15.36)与式(15.37)，只是 $\exp(\mathbf{x}'_{jl}\boldsymbol{\beta}_j/\rho_j)$ 要用 $\exp(\mathbf{x}'_{jl}\boldsymbol{\beta}_j)$ 来代替。由于 CL 是满足 $\rho_j=1$ 的式(15.36)与式(15.37)的特殊情况，所以这拥有一种可供选择的作为 CL 模型的自然推广的推导。参见麦克法登(McFadden，1978，第 79 页)、马达拉(Maddala，1983，第 70 页)以及格林(Greene，2003，第 726 页)。

非常重要的是，要注意到，当 ρ_j 对不同选项各不相同时，出现的两种变形就不一样；参见科佩尔曼和温(Koppelman and Wen，2003，第 88 页)。对嵌套 logit 模型产生怀疑，某些早期研究所获得的序贯估计本质上不同于 FIML 估计。不过，在此类研究中，各种不同估计量可应用到嵌套 logit 模型的不同变形上。此外，甚至当今各种软件包都可以估计各种变形模型。

嵌套 logit 模型能推广到较高水平的选项上(或嵌套上)。例如，戈德堡(Goldberg，1995)给出 5 个水平：(1) 购买汽车；(2) 购买给定的新车；(3) 购买已经用过 2 年的 9 类汽车的一类；(4) 外国产的车或国产车；(5) 式样。如果某个嵌套具有许多选择，那么其他引人注目之处是，足以根据选项的固定选取子集或者随机选取子集来进行估计。

15.6.5 福利分析

关于 ARUM 的福利分析，已在 15.54 节阐述过。通常，E[CV]并不存在解，即预期的补偿变化。

值得注意的是，对于收入为线性的 GEV 模型来说，$V(I-p_j,\mathbf{x}_j)=\alpha(I-p_j)+f(\mathbf{x}_j)$，麦克法登(McFadden，1995)以及早先一些研究者证明，存在显性解：

$$\mathrm{E}[CV]=\frac{1}{\alpha}(\ln G(e^{V_1''},\cdots,e^{V_m''})-\ln G(e^{V_1'},\cdots,e^{V_m'}))$$

其中,关于 GEV 分布的函数 $G(\cdot)$ 已在式(15.34)中定义,而 V_j' 与 V_j'' 表示效用的确定性成分前后的值。

不过,就含有收入的 GEV 模型而言,不存在显性解。于是,一种方法是由 15.54 节给出的模拟方法。对于多项式 logit 模型来说,这是简单的,因为很容易利用 12.8.2 节的变换方法抽取极值误差,也就是说,从(0, 1)均匀分布上采样,然后令 $\varepsilon=-\ln(-\ln(u))$。然而,对于更一般的嵌套 logit 模型来说,从 GEV 分布中进行随机采样很困难,甚至对像二变量极值那样如此简单的情况亦如此。麦克法登(McFadden, 1995)曾经提出,利用满足梅特罗波利斯—黑斯廷斯算法的 MCMC(参见 13.5 节)。赫里格斯和克林(Herringes and King, 1999)给出利用包括超越对数的各种间接效用函数,将这种模拟方法用于 15.2 节钓鱼数据的嵌套 logit 模型上的一个极好综述。

最近,达格斯文科和卡尔斯特罗姆(Dagsvik and Karlström, 2004)进一步证明,尽管若以非线性方式包括收入,GEV 模型就不会存在 E[CV]的显性解,但在解析形式上可能将 E[CV] 简化成一维积分。与利用前面提及的模拟方法相比,运用高斯积分对此积分进行计算,将更加简单一些。

15.7 随机参数 logit

随机参数 logit 模型提供一种简单方式来推广 MNL 或 CL 模型,以使每个选项的效用成为相关的。该模型或许是微观经济计量学关于横截面数据的随机参数模型的重要例子。

15.7.1 *随机参数 logit 模型*

随机参数 logit (RPL)模型[**random parameters logit (RPL) model**]是将第 i 个个体对第 j 选项的效用设定成:

$$U_{ij}=\mathbf{x}_{ij}'\boldsymbol{\beta}_i+\varepsilon_{ij},\quad j=1,2,\cdots,m \tag{15.39}$$

其中,ε_{ij} 表示 iid 的极值,如同 CL 模型一样,只是另外允许参数 $\boldsymbol{\beta}_i$ 成为随机的。一种最普遍的假设是:

$$\boldsymbol{\beta}_i\sim\mathcal{N}[\boldsymbol{\beta},\boldsymbol{\Sigma}_{\beta}] \tag{15.40}$$

一种变形是运用参数的对数正态分布,而不是正态分布,其符号为已知先验。倘若在面板背景下借用含有随机参数模型的术语,可将这个模型也称为混合 logit 模型(**mixed logit model**)。通过将 MNL 模型重新表述 CL 模型,所得到的结果同样涵盖随机参数的 MNL 模型。

此模型能重新写成:

$$U_{ij}=\mathbf{x}_{ij}'\boldsymbol{\beta}+v_{ij}$$
$$v_{ij}=\mathbf{x}_{ij}'\mathbf{u}_i+\varepsilon_{ij}$$

其中，$\mathbf{u}_i \sim \mathcal{N}[\mathbf{0}, \mathbf{\Sigma}_\beta]$。于是，$\mathrm{Cov}[v_{ij}, v_{ik}] = \mathbf{x}'_{ij}\mathbf{\Sigma}_\beta \mathbf{x}_{ik}$，$j \neq k$。因此，引进随机参数具有引人注目的性质：可推导出不同选项之间的相关性。

在大多数应用中，协方差矩阵 $\mathbf{\Sigma}_\beta$ 被设定成对角的，而且一些对角元素额外地为 0。于是，要估计的协方差参数的个数等于 $\boldsymbol{\beta}_i$ 的设定成随机的分量个数。

举一个例子，考察含有纯量回归元且参数为 β 与 σ_β^2 的混合 CL 模型。假定参数估计值是 $\hat{\beta}=2.0$，其标准误差为 0.5，同时 $\hat{\sigma}_\beta^2=1.0$，其标准误差为 0.2。于是，由于 $t=1.0/0.2=5.0$，所以常值参数的零假设即 $\hat{\sigma}_\beta^2=0$ 被强烈地拒绝。x_{ij} 增大对 $\Pr[y_i=j]$ 的效应会随着个体而变化，同时是正的且为样本的大约 97.5%，因为估计表明 $\beta_i \sim \mathcal{N}[2.0, 1.0]$。对于强调估计系数的应用来说，参见雷维尔特和特雷恩(Revelt and Train, 1998)。

行业组织文献考虑了，类似于利用市场水平(**market level data**)数据对模型用户进行加总(**aggregation**)，以此估计需求参数的 RPL 模型。例如，参见贝里(Berry, 1994)与诺夫(Nevo, 1994)，以及艾伦比和罗西(Allenby and Rossi, 1991)。

15.7.2 *随机参数 logit 的估计*

在含有随机参数的线性回归模型中，OLS 估计会产生均值 $\boldsymbol{\beta}$ 的估计，这尽管是无效的，却是一致估计值。然而，在非线性模型中，因为参数的随机性而无法控制的估计量将是非一致的。因而，如果数据生成过程是由式(15.39)与式(15.40)给出的，那么通常的条件 logit MLE 将是非一致的。然而，ML 估计必须以显性方式解释关于 $\boldsymbol{\beta}_i$ 的随机过程。

若 $\boldsymbol{\beta}_i$ 是已知的，因此唯一的随机性来源是 ε_{ij}，则以概率 $p_{ij}=e^{\mathbf{x}'_{ij}\boldsymbol{\beta}_i}/\sum_{l=1}^m e^{\mathbf{x}'_{il}\boldsymbol{\beta}_i}$ 可获得 CL 模型。实际上，由于 $\boldsymbol{\beta}_i$ 是随机的，我们要通过积分去掉该随机性。从而，得到：

$$p_{ij}=\Pr[y_i=j]=\int \frac{e^{\mathbf{x}'_{ij}\boldsymbol{\beta}_i}}{\sum_{l=1}^m e^{\mathbf{x}'_{il}\boldsymbol{\beta}_i}}\phi(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \Sigma_\beta)\,d\boldsymbol{\beta}_i \tag{15.41}$$

其中积分是多维数的，而 $\phi(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \Sigma_\beta)$ 表示关于 $\boldsymbol{\beta}_i$ 的多变量正态密度，其均值为 $\boldsymbol{\beta}$ 且方差为 Σ_β。

MLE 对 $\ln L_N=\sum_{i=1}^N\sum_{j=1}^m y_{ij}\ln p_{ij}$ 求关于 $\boldsymbol{\beta}$ 与 Σ_β 的极大值。其挑战是积分不存在闭形式解，积分维数是由 $\boldsymbol{\beta}_i$ 的分量个数给出的，可是 $\boldsymbol{\beta}_i$ 为随机的并具有非零方差。因此，通过模拟方法加以估计。

一种方法是利用直接模拟器逼近 p_{ij}(参见 12.4.1 节)。这要用被积函数在从 $\mathcal{N}[\boldsymbol{\beta}, \Sigma_\beta]$ 分布中随机采样 $\boldsymbol{\beta}_i$ 处 S 个计算值的平均值来代替积分(15.41)。于是，MSL 估计量(**MSL estimator**)为：

$$\ln \hat{\mathrm{L}}_N(\boldsymbol{\beta}, \Sigma_\beta)=\sum_{i=1}^N\sum_{j=1}^m y_{ij}\ln\left[\frac{1}{S}\sum_{s=1}^S \frac{e^{\mathbf{x}'_{ij}\boldsymbol{\beta}_i^{(s)}}}{\sum_{l=1}^m e^{\mathbf{x}'_{il}\boldsymbol{\beta}_i^{(s)}}}\right] \tag{15.42}$$

其中，$\boldsymbol{\beta}_i^{(s)}$，$s=1,\cdots,S$ 表示从密度 $\phi(\boldsymbol{\beta}_i;\boldsymbol{\beta},\Sigma_\beta)$ 获得的随机采样，由于 $\boldsymbol{\beta}$ 与 Σ_β 均是未知的，所以这种求和被嵌入在 $\boldsymbol{\beta}^{(r)}$ 与 $\Sigma_\beta^{(r)}$ 处计算的 r 次迭代程序。一致性需要 $S\to\infty$ 以及 $N\to\infty$，同时 $\sqrt{N}/S\to\infty$(参见 12.4.3 节)。快速计算方法包括使

用霍尔顿序列(参见 12.7.4 节)以及可供选择的模拟器。

一种可供选择的估计量是,运用具有相对平坦先验的贝叶斯方法。特雷恩(Train, 2001,2003)设定阶层先验满足 $\boldsymbol{\beta} \sim \mathcal{N}[\boldsymbol{\beta}^*, \boldsymbol{\Omega}^*]$,其中假定 $\boldsymbol{\Omega}^*$ 是大的,而假定 Σ_β 是逆威沙特分布,其自由度 $K=\dim[\boldsymbol{\beta}]$,而且标示度数 $\mathbf{I}_K$。为了另外包括 $\boldsymbol{\beta}_i$, $i=1,\cdots,N$,宁愿以 $\boldsymbol{\beta}$ 与 Σ_β 的后验开始研究,它在计算上比较迅速。于是有:(1) 关于 $\boldsymbol{\beta}|\Sigma_\beta$, $\boldsymbol{\beta}_i$ 的条件后验是正态的;(2) 关于 $\Sigma_\beta|\boldsymbol{\beta}$, $\boldsymbol{\beta}_i$ 的条件后验是逆威沙特的;(3) 关于 $\boldsymbol{\beta}_i|\Sigma_\beta$ 的条件后验是 $\boldsymbol{\beta}$,这与式(15.41)的被积函数成比例。已知这些条件后验,利用吉布斯抽样器的变形加以估计(参见 13.5.2 节),其新的复杂问题是,对第三个后验采样需要运用梅特罗波利斯—黑斯廷斯算法迭代(参见 13.5.4 节),因为没有完整的条件集合可以利用。在应用中,已知相对平坦先验,计算会花费掉类似于 MSL 估计量的计算时间,得到的参数估计值与标准误差通常位于源自 MSL 估计的那些值的 10%之内。

15.7.3 *广义随机效用模型*

比多项式 logit 更灵活的模型是人们所期盼的。就此而言,最近人们对随机参数 logit 模型投入了极大热忱。麦克法登和特雷恩(McFadden and Train, 2000)已经证明,任何随机效用模型都能很好地通过混合模型来任意逼近,尽管这个结果需要对回归元与混合分布进行适当选择。

把随机参数方法限制到多项式 logit 模型上并不存在什么缘由。例如,它可被推广到嵌套 logit 模型上。另外,随机性的额外来源可被并入进来,尤其是潜类型与潜变量。

为了阐述这些表达式,我们以 ARUM(15.22)来开始。这里将个体 i 对第 j 个选项的效用设定成 $U_{ij}=V_{ij}(\mathbf{x}_i,\boldsymbol{\beta})+\varepsilon_{ij}$,其中,$\mathbf{x}_i$ 表示观测数据,$\boldsymbol{\beta}$ 表示未知参数,而 ε_{ij} 表示误差,对于不同 i 来说 ε_{ij} 是独立的,但对不同 j 而言 ε_{ij} 可能是相关的。假定 ε_{ij} 的分布使得式(15.23)产生选择概率的闭形式解,该选择概率记为:

$$p_{ij}=F_j(\mathbf{V}_i(\mathbf{x}_i,\boldsymbol{\beta}),\boldsymbol{\theta}_\varepsilon)$$

其中,$\mathbf{V}_i(\mathbf{x}_i,\boldsymbol{\beta})=[V_{i1}(\mathbf{x}_i,\boldsymbol{\beta}),\cdots,V_{im}(\mathbf{x}_i,\boldsymbol{\beta})]$,而 $\boldsymbol{\theta}_\varepsilon$ 表示 $\varepsilon_i=(\varepsilon_{i1},\cdots,\varepsilon_{im})$ 分布的任何未知参数。倘若 ε_i 服从 GEV 分布,可能得出这种闭形式解,对于特定情况来说,会导致多项式 logit 模型以及嵌套 logit 模型。

更一般的模型将引入其他的随机性。首先,前面的效用确定部分变成 $V_{ij}=V_{ij}(\mathbf{x}_i, \boldsymbol{\xi}_i, \boldsymbol{\beta})$ 。然后,假定 ε_i 使得概率的闭形式解是以 $\boldsymbol{\xi}_i$ 为条件而存在,无条件出现:

$$p_{ij}=\int F_j(\mathbf{V}_i(\mathbf{x}_i, \boldsymbol{\xi}_i, \boldsymbol{\beta}), \boldsymbol{\theta}_\varepsilon)f(\boldsymbol{\xi}_i|\boldsymbol{\theta}_\xi)d\boldsymbol{\xi}_i \qquad (15.43)$$

其中,$f(\boldsymbol{\xi}|\boldsymbol{\theta}_\xi)$表示 $\boldsymbol{\xi}$ 的密度。RPL 模型是满足 $V_{ij}=\mathbf{x}_{ij}'\boldsymbol{\beta}+\mathbf{x}_{ij}'\boldsymbol{\xi}_i$ 的一个例子,其中,$\boldsymbol{\xi}_i$ 服从 $\mathcal{N}[\mathbf{0}, \boldsymbol{\Sigma}]$,同时经由随机参数自变量而激发出来。不过,还可引进 $\boldsymbol{\xi}_i$ 作为另外分布项或有关的潜变量。其次,假定个体者是来自 C 潜类型的一个;参见 18.5 节关于持续期限模型的例子,以及斯威特(Swait, 2003)关于潜类型 GEV 例

子或有限混合模型。若 $\boldsymbol{\beta}$ 与 $\boldsymbol{\theta}_\varepsilon$ 通过类型而变化，则式(15.43)无条件变成：

$$p_{ij}=\sum_{c=1}^{C}\left[\int F_j(\mathbf{V}_i(\mathbf{x}_i,\boldsymbol{\xi}_i,\boldsymbol{\beta}^c),\boldsymbol{\theta}_\varepsilon^c)f(\boldsymbol{\xi}_i|\boldsymbol{\theta}_\xi)d\boldsymbol{\xi}_i\right]\pi_c \tag{15.44}$$

其中，π_c 表示作为第 c 类型成员资格的概率，而且典型地有 $c=2$ 或 $c=3$。于是，MSL 估计量对：

$$\ln\hat{\mathrm{L}}_N(\boldsymbol{\beta},\Sigma_{\boldsymbol{\beta}})=\sum_{i=1}^{N}\sum_{j=1}^{m}y_{ij}\ln\left[\frac{1}{S}\sum_{s=1}^{S}\sum_{c=1}^{C}F_j(\mathbf{V}_i(\mathbf{x}_i,\boldsymbol{\xi}_i^s,\boldsymbol{\beta}^c),\boldsymbol{\theta}_\varepsilon^c)\pi_c\right]$$

求极大值，其中，$\boldsymbol{\xi}_i^s$ 表示从 $f(\boldsymbol{\xi}_i^s|\boldsymbol{\theta}_\xi)$ 中得到的第 s 个采样。卡马库兰和韦德尔(Kamakura and Wedel, 2004)曾经利用贝叶斯方法估计有限混合 MNL 模型。

沃克和本·阿基瓦(Walker and Ben-Akiva, 2002)将这类模型称为广义随机效用模型(**generalized random utility model**)。他们引用许多文章来进行此类推广，考虑使用意向偏好数据补充显性偏好数据(**stated preference data**)，同时提供内容充实的实证说明。源自沃克和本·阿基瓦(Walker and Ben-Akiva, 2002)的图15.1概括出各种扩展。

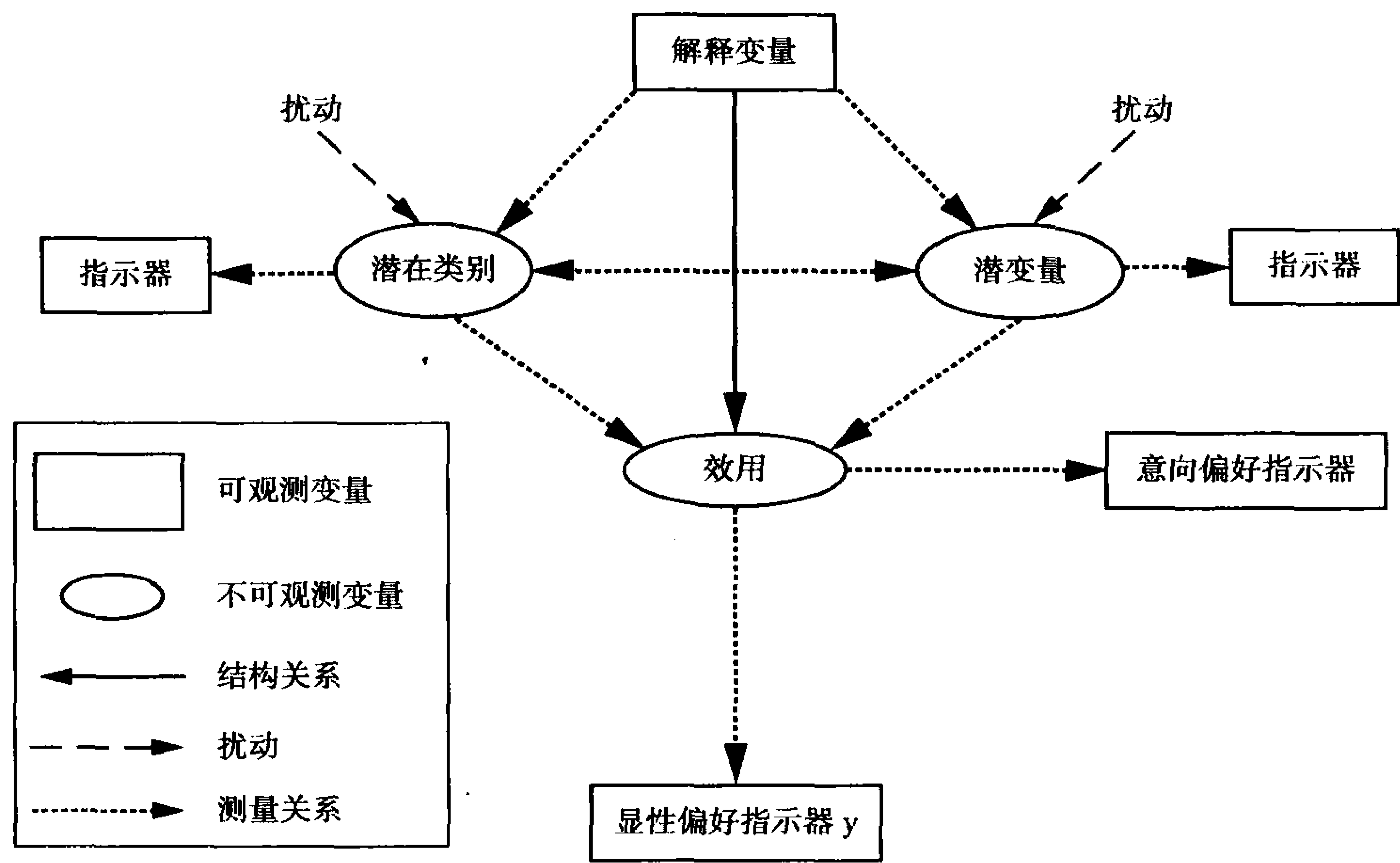


图 15.1 广义随机效用模型

多项式建模文献处于发展与估计高度有结构的参数模型的最前沿，此参数模型并入了随机参数、潜变量、潜参数，并且将一个以上来源的数据组合起来。这些方法可用于任何类型的横截面数据，而不只是用于离散结果。

15.8 多项式 probit

一种可供选择的引入不可观测成分中关于不同选择具有相关性且明显的方式是，以正态分布误差研究。不过，很难进行 ML 估计，如同最一般情况一样，需要计

算$(m-1)$重积分。

15.8.1 多项式 probit 模型

多项式 probit(MNP)模型是 m 个选择的多项式模型,其第 j 个选择的效用为:

$$U_j=V_j+\varepsilon_j,\quad j=1,2,\cdots,m \tag{15.45}$$

其中,误差服从联合正态分布,满足:

$$\varepsilon \sim \mathcal{N}[\mathbf{0},\ \mathbf{\Sigma}] \tag{15.46}$$

这里,$m\times1$ 维向量 $\varepsilon_1=[\varepsilon_1,\cdots,\varepsilon_m]'$。通常 $V_j=\mathbf{x}_j'\beta$ 或 $V_j=\mathbf{x}'\beta_j$。

各种不同 MNP 模型起因于对协方差矩阵 Σ 的不同设定。非对角线的一些元素被设定为非零的,以便允许对不同误差具有相关性,尽管需要对 Σ 施加某些限制。注意到,如果误差是不相关的,那么 MNP 仍不会产生概率的闭形式解,然而,比较容易的是,假定误差是极值的且运用 CL 模型或 MNL 模型。

为了确保识别(**identification**),需要对 Σ 加以限制。很明显,由式(15.23)知,对于任何 ARUM 来说,选择可由效用或误差之差来确定。因而,将选取作为基准选项之后,我们考察选项 j 的效用与选项 1 的效用之差。邦奇(Bunch, 1991)曾经证明,除了误差 $\varepsilon_j-\varepsilon_1$ 的协方差矩阵的一个参数之外,全部是可识别的。参见 15.5.1 节结尾的讨论。达到这种识别的一种方法是正规化,比如说 $\varepsilon_1=0$,然后对协方差元素限制成 1。例如,当 $m=2$ 时,则设 $\varepsilon_1=0$,所以 $\sigma_{11}=0$ 且 $\sigma_{12}=0$,并且另外限制 $\sigma_{22}=1$。于是,$\varepsilon_2-\varepsilon_1=\varepsilon_2\sim\mathcal{N}[0,1]$,这是一个二值 probit 模型。

为了成功应用,需要对 Σ 或 β进行额外限制。基恩(Keane, 1992)已经证明,即使为了确保恰好识别而对误差协方差做出一些假设,实际上在含有不随选项而变化的回归元的模型中,MNP 模型的参数可能是非常不精确的估计。这种估计不精确在性质上类似于线性回归中回归元之间的高度多重共线性。基恩发现,关于回归元排除性约束会很好地发挥作用(对每个效用指标具有一个排除性约束)。作为一种可供选择的且更普遍的方式是,对协方差参数施加进一步限制。

关于误差的一种流行而简约模型是因子模型[1](**factor model**)

$$\varepsilon_j=v_j+\sum_{l=1}^{L}c_{jl}\xi_l,\quad j=1,2,\cdots,m$$

其中,v_j 与 $\xi_1,\cdots,\xi_L$ 都是服从 iid 标准正态的,c_{jl}表示权重,c_{jl}被称为待估因子载荷(**factor loadings**)。该模型能极大地将协方差参数的数目从 $m(m+1)/2$ 减少到 L,并需要一个$(L+1)$维的积分。对于小 L 值来说,可运用数值方法,通常是高斯积分,而对于大 L 值来说,则需要使用模拟方法。就面板数据而言,将随机效应模型看成是含有误差 $u_{it}=\alpha_i+\epsilon_{it}$ 的因素模型,而因子模型尤其适合于面板 probit 背景下的情况。

---

[1] 也称为因素模型。——译者注

15.8.2 多项式 probit 的估计

回归以及误差方差参数可更好地通过 15.3.2 节给出的对数似然 ML 来进行估计。其挑战是，选择概率的表达式并不存在闭形式解。

对于三种选择 MNP 模型来说，有：

$$p_1 = \Pr[y=1] = \int_{-\infty}^{-\tilde{V}_{31}} \int_{-\infty}^{-\tilde{V}_{21}} f(\tilde{\varepsilon}_{21}, \tilde{\varepsilon}_{31}) d\tilde{\varepsilon}_{21} d\tilde{\varepsilon}_{31}$$

[参见式(15.24)]，其中，$f(\tilde{\varepsilon}_{21}, \tilde{\varepsilon}_{31})$表示具有两个自由二变量正态的协方差参数，而 $\tilde{V}_{21}$与 $\tilde{V}_{31}$均依赖于回归元与参数 $\boldsymbol{\beta}$。这个二变量正态积分能在数值形式上迅速地计算出来。然而，更一般地，m 个选择的模型需要数值计算$(m-1)$个变量积分。将标准数值积分方法限定在四种选择 MNP 模型上，三变量正态积分变成数值方法的极限。

对于较大的模型来说，一种可供选择的方法是使用模拟方法。为了简单起见，我们涉及三种选择的 MNP 模型。一种可能性是使用频率模拟器，即通过对小于$(-\tilde{V}_{21}, -\tilde{V}_{31})$的抽取$(\tilde{\varepsilon}_{21}, \tilde{\varepsilon}_{31})$的部分来逼近 p_1。由 12.7.1 节知，这一模拟器不是光滑的，并且它可以是非常无效的(参见 12.7.2 节)。进一步地，在当前背景下，可能情况是，它会得到 $\hat{p}_1=0$ 或 1 的边界值。通常来说，一种较好方式是运用重要抽样，详细内容由 12.7.2 节阐述。就多变量正态区域上进行蒙特卡罗积分而言，一种极为流行的重要抽样器是 GHK 模拟器，该方法归功于格韦克(Grweke, 1992)、哈吉瓦西柳和麦克法登(Hajivassiliou and McFadden, 1994)，以及基恩(Keane, 1994)。这会递推地截取多变量正态 pdf。和频率模拟器相比，它是光滑的，对于选项需要较少的以小概率进行的抽取，同时不可能具有边界问题。特雷恩(Train, 2003)对该方法提供了详细解释。

前面讨论考察了，假定知道 $\boldsymbol{\beta}$ 与 Σ 来计算 MNP 概率的问题。实际上，我们需要估计 $\boldsymbol{\beta}$ 与 Σ。模拟极大似然估计量(**maximum simulated likelihood estimator**)极大化：

$$\ln \hat{\mathrm{L}}_N(\boldsymbol{\beta}, \Sigma) = \sum_{i=1}^{N} \sum_{j=1}^{m} y_{ij} \ln \hat{p}_{ij}$$

其中，$\hat{p}_{ij}$表示利用 GHK 或其他估计量而获得的。一致性要求模拟器中的采样数量 $S \to \infty$以及 $N \to \infty$ 。该方法显得相当累赘。在迭代进行第 r 次时(参见第 10 章)，估计值是 $\hat{\boldsymbol{\beta}}^{(r)}$ 与 $\hat{\Sigma}^{(r)}$，同时更新需要重新计算 $\hat{p}_{ij}$，这要求对 N 个个体中的每一个都要进行 S 个采样。

一种可供选择的估计方法是模拟矩方法(**method of simulated moments**)(参见 12.5 节)。由式(15.8)知，一致矩方法估计量是$\sum_{i=1}^{N}\sum_{j=1}^{m}(y_{ij}-p_{ij})\mathbf{z}_i=\mathbf{0}$ 的解，其中，例如 $\mathbf{z}_i=\mathbf{x}_i$。相对应的 $\boldsymbol{\beta}$ 与 Σ MSM 估计量是估计方程

$$\sum_{i=1}^{N} \sum_{j=1}^{m} (y_{ij} - \hat{p}_{ij}) \mathbf{z}_i = \mathbf{0}$$

的解，其中，$\hat{p}_{ij}$是利用无偏模拟器获得的。于是，$(y_{ij}-\hat{p}_{ij})\mathbf{z}_i$ 关于$(y_{ij}-p_{ij})\mathbf{z}_i$ 是无偏的，因此，即使 $S=1$，也可能是一致估计的。这大大简化了计算。不过，就小 S

而言，有效性出现损失，而且甚至对大 S 而言，与 MSL 相比，MSM 有效性更差一些，因为在本例中，矩方法的有效性比 ML 的要差一些。与 MSL 同样有效的极少运用的方法是模拟得分方法(**method of simulated scores**)[参见哈吉瓦西柳和麦克法登(Hajivassiliou and McFadden, 1998)]。

一种可供选择的估计量是运用贝叶斯方法。与 RPL 不同，概率不存在闭形式解，这需要从效用中推导出来。引进潜效用 $\mathbf{U}_i=(U_{1i},\cdots,U_{ji})$ 作为辅助变量，并运用数据增广方法(参见 13.7 节)。若令 $\mathbf{U}=(U_1,\cdots,U_N)$ 且 $\mathbf{y}=(y_1,\cdots,y_N)$，我们会使得吉布斯抽样器在：(1) 关于 $\boldsymbol{\beta}|\mathbf{y},\mathbf{U},\Sigma$ 的条件后验；(2) 关于 $\Sigma|\mathbf{y},\boldsymbol{\beta},\mathbf{U}$ 的条件后验；(3) 关于 $\mathbf{U}_i|\mathbf{y},\boldsymbol{\beta},\Sigma$ 的后验之间进行循环。艾伯特和奇布提供既有无序多项式模型，又有有序多项式模型的相当一般的研究。麦卡洛克和罗西(McCulloch and Rossi, 1994)曾提供了内容丰富的 MNP 应用。奇布(Chib, 2001)已经讨论过为了识别需要利用 Σ 约束的复杂情况(参见 15.8.1 节)。

15.8.3 *讨论*

MNP 模型既缺乏 p_{ij} 的闭形式解，RPL 模型也缺乏 $\boldsymbol{\beta}_i$ 的闭形式解。不过，就 RPL 而言，至少存在以 $\boldsymbol{\beta}_i$ 为条件的闭形式解，且唯一问题是通过积分去掉 $\boldsymbol{\beta}_i$。对于 MNP 模型来说，它在时间上先于 RPL 模型，尤其是当 p_{ij} 接近于 0 或 1 时，没有此类条件结果而且逼近 p_{ij} 则是更富有挑战性的。看起来，通过嵌套 logit、RPL 或混合模型而不是使用 MNP，更容易获得模型灵活性。

15.9 有序、序列和分级结果

在本节，我们阐述比无序模型更具有结构性的模型，诸如那些含有自然顺序的选项或者依次决策的模型。当很容易建立起合适模型时，可直接进行分析，而且再次利用建立在式(15.4)上的 MLE 加以估计，各种不同的模型会导致对概率 p_{ij} 的不同设定。

15.9.1 *有序多项式模型*

假定选项存在一种自然顺序。例如，健康自我评价状况可以是极好、良好、一般或不好。这类数据能通过无序多项式模型估计，但更为简约的模型以及切合实际的模型是要将这种顺序考虑进去的。

起点是含有单个潜变量的指标模型：

$$y_i^*=\mathbf{x}_i'\boldsymbol{\beta}+u_i \tag{15.47}$$

其中，$\mathbf{x}$ 并不包括截距，这违背了 14.4.1 节内容。由于 y^* 跨越一系列递增的未知门限值，我们就往上移动选项的次序。例如，对于非常小的 y^*，健康状况是不好的；对于 $y^*>\alpha_1$，健康状况改进到一般；对于 $y^*>\alpha_2$，健康进一步改善到良好等。

通常，就 m 个选项的有序模型而言，定义：

$$当\ \alpha_{j-1}<y_i^*\leqslant\alpha_j\ 时，\ y_i=j \tag{15.48}$$

其中 $\alpha_0=-\infty$ 以及 $\alpha_m=\infty$，于是：

$$\begin{aligned}\Pr[y_i=j]&=\Pr[\alpha_{j-1}<y_i^*\leqslant\alpha_j]\\&=\Pr[\alpha_{j-1}<\mathbf{x}_i'\boldsymbol{\beta}+u_i\leqslant\alpha_j]\\&=\Pr[\alpha_{j-1}-\mathbf{x}_i'\boldsymbol{\beta}<u_i\leqslant\alpha_j-\mathbf{x}_i'\boldsymbol{\beta}]\\&=F(\alpha_j-\mathbf{x}_i'\boldsymbol{\beta})-F(\alpha_{j-1}-\mathbf{x}_i'\boldsymbol{\beta})\end{aligned}\tag{15.49}$$

其中，F 表示 u_i 的 cdf。回归参数 $\boldsymbol{\beta}$ 与 $(m-1)$ 个门限参数 $\alpha_1,\cdots,\alpha_{m-1}$ 均可通过对含有式(15.49)定义的 p_{ij} 对数似然(15.5)求极大值而获得。对于有序 logit 模型(**order logit model**)来说，u 表示满足 $F(z)=e^z/(1+e^z)$ 的逻辑斯蒂分布。对于有序 probit 模型(**order probit model**)来说，u 表示标准正态分布，而且 $F(\cdot)$ 表示标准正态的 cdf。令 K 表示把截距排除在外的回归元数量，m 个选项的有序模型具有 $K+m-1$ 个参数，而 MNL 模型则具有 $(m-1)(K+1)$ 个参数。

对回归参数 $\boldsymbol{\beta}$ 的符号，可立刻解释成确定潜变量 y^* 是否随回归元而增大。就概率边际效应而言，有：

$$\frac{\partial\Pr[y_i=j]}{\partial\mathbf{x}_i}=\{F'(\alpha_{j-1}-\mathbf{x}_i'\boldsymbol{\beta})-F'(\alpha_j-\mathbf{x}_i'\boldsymbol{\beta})\}\boldsymbol{\beta}$$

其中，F' 表示 F 的导数。括号中的项可正可负。

这种模型还能用于仅取几个值的计数数据。卡梅伦和特里维迪(Cameron and Trivedi, 1986)将有序 probit 模型用于医生会诊次数。豪斯曼、洛和麦金利(Hausman, Lo, and MacKinley, 1992)则把有序 probit 用于计数变动数据，这可以是负的，此外，可将误差项 u_i 建模成异方差的。

15.9.2 *序列多项式模型*

在一些情况下，决策要求序贯做出。例如，人们首先决定是否去上大学。如果选择不上大学，那么 $y=1$。如果 $y\neq1$，那么决定是否上两年制大学($y=2$)或四年制大学($y=3$)。给定此序列的设定，很容易获得其概率。例如，对第一次决策通过 probit 模型来建模，并且如果有意义，对第二次决策仍通过 probit 建模。于是，$\Pr[y=1]=\Phi(\mathbf{x}_1'\boldsymbol{\beta}_1)$ 以及 $\Pr[y=2|y\neq1]=\Phi(\mathbf{x}_2'\boldsymbol{\beta}_2)$。其无条件概率是：

$$\Pr[y=2]=\Pr[y=2|y\neq1]\times\Pr[y\neq1]=\Phi(\mathbf{x}_2'\boldsymbol{\beta}_2)(1-\Phi(\mathbf{x}_1'\boldsymbol{\beta}_1))$$

参数 $\boldsymbol{\beta}_1$ 与 $\boldsymbol{\beta}_2$ 可通过对数似然函数(15.5)求极大值而估计出来，其中，$p_{1i}=\Phi(\mathbf{x}_{1i}'\boldsymbol{\beta}_1)$，$p_{2i}$ 已由前面方程给出，而 $p_{3i}=1-p_{1i}-p_{21}$。

这种方法依赖于对做出决策序列进行正确设定。就此选择例子而言，一个较好的模型是三种选择嵌套的 logit 模型，其中上两年制大学效用的误差是独立的。利用由 8.5 节给出的基于似然方法，可对这些模型加以比较。

15.9.3 *分级数据模型*

因而，对于假定选项是不相容的且唯一选项被选取的模型已经进行了讨论。更一般地，选项是分等级的，尤其是还有意向偏好数据的情形。例如，已知第一个选项和第二个选项。

可直接进行估计分级有序 logit 模型(**rank-ordered logit model**)[参见贝格斯、卡德尔和豪斯曼(Beggs, Cardell, and Hausman, 1981)]。考察四个选项条件 logit 模型,其具有选项 2 作为第一种选择,而具有选项 3 作为第二种选择。选项 2 是从所有四个选项中选取的,然后选项 3 从剩余的选项 1、选项 3、选项 4 这三个选项中选取。第一种选择与第二种选择的联合概率是:

$$\frac{e^{\mathbf{x}_{i2}'\beta}}{e^{\mathbf{x}_{i1}'\beta}+e^{\mathbf{x}_{i2}'\beta}+e^{\mathbf{x}_{i3}'\beta}+e^{\mathbf{x}_{i4}'\beta}}\times\frac{e^{\mathbf{x}_{i3}'\beta}}{e^{\mathbf{x}_{i1}'\beta}+e^{\mathbf{x}_{i3}'\beta}+e^{\mathbf{x}_{i4}'\beta}}$$

给定关于其他 11 个联合概率的类似表达式,可通过 ML 进行估计。

对于多项式 probit 模型来说,不存在类似简化。哈吉瓦西柳和鲁德(Hajivassiliou and Ruud, 1994)曾阐述了模拟联合概率的方法。他们运用分级有序 probit 模型(**rank-ordered probit model**)阐明各种基于模拟的估计量。

15.10 多变量离散结果

前面一些模型,除分级有序模型之外,都是单一离散因变量在 m 个互不相交值中取一个。现在,我们考察存在一个以上离散结果的模型。对数似然函数类似于多项式模型(15.5),只是各种不同模型对应于概率不同的函数形式。为了解释两个结果之间的相关以及可能的联立性,需要这些概率。

15.10.1 *二变量离散结果*

为了简单起见,考察二变量离散数据(**bivariate discrete data**)(y_{1i},y_{2i})。例如,在劳动力供给与生育率的联合模型中,关于个体 i 的因变量(y_{1i},y_{2i})可能是若工作则 $y_{1i}=2$,若不工作则 $y_{1i}=1$;若有小孩则 $y_{2i}=2$,若没有小孩则 $y_{2i}=1$。

更一般地讲,y_1 可取值 $1,\cdots,m_1$, 而 y_2 可取值 $1,\cdots,m_2$。对于个体 i 来说,定义:

$$p_{ijk}=\Pr[y_{1i}=j,y_{2i}=k],\quad j=1,\cdots,m_1,\ k=1,\cdots,m_2 \tag{15.50}$$

注意到,p_{ijk}定义了互不相交事件的概率,并且$\sum_j\sum_k p_{ijk}=1$。定义 $m_1\times m_2$ 对应于二值指示变量,若$(y_1=j,\ y_2=k)$,则 $y_{jk}=1$,否则 $y_{jk}=0$。于是,第 i 个观测值的联合密度是:

$$f(y_{1i},y_{2i})=\prod_{k=1}^{m_1}\prod_{j=1}^{m_2}p_{ijk}^{y_{ijk}}$$

从而,对数似然是$\sum_{i=1}^{N}\sum_{k=1}^{m_1}\sum_{j=1}^{m_2}y_{ijk}\ln p_{ijk}$,并如同 15.4.2 节一样,通过 ML 加以估计。

在多变量模型与多项式模型之间,其本质差别在于对概率函数形式的设定上。

在最简单情况下,两个离散因变量是独立的,且 $p_{ijk}=\Pr[y_{1i}=j]\times\Pr[y_{2i}=k]$。于是,$y_1$ 与 y_2 能利用各自的多项式模型加以建模。

不过,当将两个变量看成是相互联系的,一种简单方法是使用概率 p_{ijk}的多项式 logit 模型。从而,本质上可将二变量结果(y_1,y_2)看成 $m_1\times m_2$ 个单变量结果。

例如,在劳动力供给与生育率的例子中,四种结果之一就是工作与有小孩。

在下一节,考察这两种极端之间的模型。

15.10.2 二变量 probit

二变量 probit 模型是关于两个二值结果的联合模型,它可推广到从一个潜变量到两个可能相关的潜变量的指标函数形式(参见 14.4.1 节)。

定义不可观测的潜变量:

$$y_1^* = \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1 \tag{15.51}$$
$$y_2^* = \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon_2$$

其中,ε_1 与 ε_2 服从联合正态分布,其均值为 0,方差为 1,且相关系数为 ρ。对于二变量 probit 模型(**bivariate probit model**),设定观测结果为:

$$y_1 = \begin{cases} 2, & \text{当 } y_1^* > 0 \\ 1, & \text{当 } y_1^* \leqslant 0 \end{cases}$$

$$y_2 = \begin{cases} 2, & \text{当 } y_2^* > 0 \\ 1, & \text{当 } y_2^* \leqslant 0 \end{cases}$$

其中,我们使用(2,1)值而不是(1,0)值,这与本章记号一致。当误差相关系数 $\rho=0$ 时,该模型就会变成关于 y_1 与 y_2 的两个单独 probit 模型。

当 $\rho \neq 0$ 时,概率不存在闭形式解。例如:

$$\begin{aligned} p_{22} &= \Pr[y_1 = 2,\ y_2 = 2] \\ &= \Pr[y_1^* > 0,\ y_2^* > 0] \\ &= \Pr[-\varepsilon_1 < \mathbf{x}_1'\boldsymbol{\beta}_1,\ -\varepsilon_2 < \mathbf{x}_2'\boldsymbol{\beta}_2] \\ &= \Pr[\varepsilon_1 < \mathbf{x}_1'\boldsymbol{\beta}_1,\ \varepsilon_2 < \mathbf{x}_2'\boldsymbol{\beta}_2] \\ &= \int_{-\infty}^{\mathbf{x}_1'\boldsymbol{\beta}_1}\int_{-\infty}^{\mathbf{x}_2'\boldsymbol{\beta}_2} \phi(z_1,\ z_2,\ \rho)dz_1dz_2 \\ &= \Phi(\mathbf{x}_1'\boldsymbol{\beta}_1,\ \mathbf{x}_2'\boldsymbol{\beta}_2,\ \rho) \end{aligned}$$

其中,$\phi(z_1,z_2,\rho)$ 与 $\Phi(z_1,z_2,\rho)$ 分别表示标准化二变量正态密度与关于 (z_1,z_2) 的 cdf,具有零均值、单位方差以及相关系数 ρ,同时对于具有零均值的二变量正态分布来说,第四个等式成立。

就其他可能结果而言,经过类似代数运算,得到:

$$\begin{aligned} p_{jk} &= \Pr[y_1 = j,\ y_2 = k] \\ &= \Phi(q_1\mathbf{x}_1'\boldsymbol{\beta}_1,\ q_2\mathbf{x}_2'\boldsymbol{\beta}_2,\ \rho) \end{aligned}$$

其中,当 $y_l=2$ 时,$q_l=1$;当 $y_l=1$ 时,$q_l=-1$,这里 $l=1,2$。这是 ML 估计的基础,对此格林(Greene, 2003)曾经详细阐述过,他还考察了边际效应计算。

具体实施需要对二变量正态积分进行计算,这样做在数值形式上是可行的。尽管由于存在较高阶积分,遇到数值计算上的挑战,但对多变量 probit 的推广是显

而易见的。如果每一个结果都是有序的，那么此模型能被推广到二变量有序 probit 模型（**bivariate ordered probit model**）。

人们还可考察推广式（15.5）的联立方程 probit 模型，以使右边变量成为内生的，例如，关于 y_1^* 的第一个方程包括 y_2^*，同时（或者）y_2 作为回归元，而且关于 y_2^* 可做出类似讨论，只是为了确保模型是可识别的而需要某些约束。这种模型类似于 16.8.2 节将要讨论的联立方程 Tobit 模型。

15.11 半参数估计

某些研究可推广到为了对无序多项式数据进行建模的半参数估计方法上。阿贝（Abe，1999）曾经估计下述 logit 模型，即用其他模型形式 $\sum_p \beta_p f_p(\mathbf{x}_{ijp})$ 代替式（15.10）中的 $\mathbf{x}_{ij}'\boldsymbol{\beta}$，其中，$p$ 表示 $\mathbf{x}_{ij}$ 的第 p 个分量，而函数 $f_p(\cdot)$ 是通过数据估计出。李龙飞（L-F. Lee，1995）将克莱因和斯帕迪（Klein and Spady，1993）的源自二值结果的估计量（参见 14.7 节）推广到多项式结果。多重指标模型的半参数方法，同样可应用于多项式无序模型。其挑战是确保预测概率位于 0 与 1 之间且和为 1。

有序模型可以很好地协助半参数分析，因为它们涉及跨越一系列门限的指标 $\mathbf{x}'\boldsymbol{\beta}$。例如，参见克莱因和舍曼（Klein and Sherman，2002），他们在误差与回归元是独立的假设下，阐述作为既关于回归又关于至位置及标度的门限点的 $\sqrt{N}$ 一致且服从渐近正态的估计量。

15.12 MNL、CL 以及 NL 模型推导

我们考察条件 logit 模型和多项式 logit 模型，推导对数似然函数的一阶导数与二阶导数，以及回归元变化对概率效应的表达式。然后，从 GEV 模型推导嵌套 logit (NL)模型。

15.12.1 条件 logit

条件 logit 概率是 $p_{ij}=e^{\mathbf{x}_{ij}'\boldsymbol{\beta}}/e^{\mathbf{x}_{il}'\boldsymbol{\beta}}$。运用分部微分，得出：

$$\frac{\partial p_{ij}}{\partial \boldsymbol{\beta}}=\frac{e^{\mathbf{x}_{ij}'\boldsymbol{\beta}}}{\sum_l e^{\mathbf{x}_{il}'\boldsymbol{\beta}}}\mathbf{x}_{ij}-\frac{e^{\mathbf{x}_{ij}'\boldsymbol{\beta}}}{(\sum_l e^{\mathbf{x}_{il}'\boldsymbol{\beta}})^2}\sum_l e^{\mathbf{x}_{i2}'\boldsymbol{\beta}}\mathbf{x}_{il}$$

$$=p_{ij}\mathbf{x}_{ij}-p_{ij}\sum_l p_{il}\mathbf{x}_{il}=p_{ij}\mathbf{x}_{ij}-p_{ij}\bar{\mathbf{x}}_i=p_{ij}(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)$$

其中，$\bar{\mathbf{x}}_i=\sum_l p_{il}\mathbf{x}_{il}$。于是：

$$\frac{\partial \mathcal{L}}{\partial \beta}=\sum_i\sum_j \frac{y_{ij}}{p_{ij}}\frac{\partial p_{ij}}{\partial \boldsymbol{\beta}}=\sum_i\sum_j \frac{y_{ij}}{p_{ij}}p_{ij}(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)=\sum_i\sum_j y_{ij}(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)$$

由此可得：

$$\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = -\sum_i \sum_j y_{ij} \frac{\partial \bar{\mathbf{x}}_i}{\partial \boldsymbol{\beta}'}$$
$$= -\sum_i \sum_j y_{ij} \frac{\partial \sum_l p_{il} \mathbf{x}_{il}}{\partial \boldsymbol{\beta}'}$$
$$= -\sum_i \sum_j y_{ij} \sum_l p_{il} (\mathbf{x}_{il} - \bar{\mathbf{x}}_i) \mathbf{x}'_{il}$$
$$= -\sum_i \sum_j p_{ij} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i) \mathbf{x}'_{ij}$$
$$= -\sum_i \sum_j p_{ij} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)'$$

它就是式(15.15)。倒数第二个等式运用了下面事实：对于恰好一个选择来说，y_{ij} 等于 1，否则 y_{ij} 为 0，因此$\sum_j y_{ij} \sum_l a_{il} = \sum_j \sum_l y_{ij} a_{il} = \sum_j a_{ij}$，同时最后一个等式运用$\sum_j p_{ij}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)\bar{\mathbf{x}}'_i = \sum_j (p_{ij}\mathbf{x}_{ij} - p_{ij}\bar{\mathbf{x}}_i)\bar{\mathbf{x}}'_i = \sum_j (\bar{\mathbf{x}}_i - p_{ij}\bar{\mathbf{x}}_i)\bar{\mathbf{x}}'_i = \mathbf{0}$，因为$\sum_j p_{ij} = 1$。

现在，考察回归元变化的效应。对于条件 logit 模型来说，有：

$$\frac{\partial p_{ij}}{\partial \mathbf{x}_{ij}} = \frac{e^{\mathbf{x}'_{ij}\boldsymbol{\beta}}}{\sum_l e^{\mathbf{x}'_{il}\boldsymbol{\beta}}} \boldsymbol{\beta} - \frac{e^{\mathbf{x}'_{ij}\boldsymbol{\beta}}}{(\sum_l e^{\mathbf{x}'_{il}\boldsymbol{\beta}})^2} e^{\mathbf{x}'_{ij}\boldsymbol{\beta}} \boldsymbol{\beta} = p_{ij}(1 - p_{ij})\boldsymbol{\beta}$$

不过，当 $j \neq k$ 时，有：

$$\frac{\partial p_{ij}}{\partial \mathbf{x}_{ik}} = -\frac{e^{\mathbf{x}'_{ij}\boldsymbol{\beta}}}{(\sum_l e^{\mathbf{x}'_{il}\boldsymbol{\beta}})^2} e^{\mathbf{x}'_{ik}\boldsymbol{\beta}} \boldsymbol{\beta} = -p_{ij} p_{ik} \boldsymbol{\beta}$$

对上述两个结果组合，得到式(15.18)。

15.12.2 多项式 logit

多项式 logit 概率是 $p_{ij} = e^{\mathbf{x}'_i\boldsymbol{\beta}_j} / \sum_l e^{\mathbf{x}'_i\boldsymbol{\beta}_l}$。通过分部微分得到：

$$\frac{\partial p_{ij}}{\partial \boldsymbol{\beta}_j} = \frac{e^{\mathbf{x}'_i\boldsymbol{\beta}_j}}{\sum_l e^{\mathbf{x}'_i\boldsymbol{\beta}_l}} \mathbf{x}_i - \frac{e^{\mathbf{x}'_i\boldsymbol{\beta}_j}}{(\sum_l e^{\mathbf{x}'_i\boldsymbol{\beta}_l})^2} e^{\mathbf{x}'_i\boldsymbol{\beta}_j} \mathbf{x}_i = p_{ij}\mathbf{x}_i - p_{ij}p_{ij}\mathbf{x}_i$$

不过，对于 $k \neq j$，有：

$$\frac{\partial p_{ij}}{\partial \boldsymbol{\beta}_k} = -\frac{e^{\mathbf{x}'_i\boldsymbol{\beta}_j}}{(\sum_l e^{\mathbf{x}'_i\boldsymbol{\beta}_l})^2} e^{\mathbf{x}'_i\boldsymbol{\beta}_k} \mathbf{x}_i = - p_{ij} p_{ik} \mathbf{x}_i$$

对上述式子组合，得出：

$$\frac{\partial p_{ij}}{\partial \boldsymbol{\beta}_k} = \delta_{ijk} p_{ij} \mathbf{x}_i - p_{ij} p_{ik} \mathbf{x}_i = p_{ij}(\delta_{ijk} - p_{ik})\mathbf{x}_i$$

其中，指示变量 $\delta_{ijk} = 1$，当 $j = k$，同时：

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}_k} = \sum_i \sum_j \frac{y_{ij}}{p_{ij}} \frac{\partial p_{ij}}{\partial \boldsymbol{\beta}_k}$$
$$= \sum_i \sum_j \frac{y_{ij}}{p_{ij}} (\delta_{ijk} p_{ij} - p_{ij} p_{ik} \mathbf{x}_i)$$
$$= \sum_i \left[\sum_j y_{ij}\delta_{ijk} - y_{ij} p_{ik} \right] \mathbf{x}_i$$
$$= \sum_i [y_{ik} - p_{ik}] \mathbf{x}_i$$

如同式(15.16)所表述的，其中最后一行运用了 δ_{ijk} 的定义以及$\sum_j y_{ij} = 1$。对于二

阶导数来说，得出：

$$\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_k'} = -\sum_i \sum_j \frac{\partial p_{ij}}{\partial \boldsymbol{\beta}_k'} \mathbf{x}_i = -\sum_i \sum_j p_{ij}(\delta_{ijk} - p_{ik}) \mathbf{x}_i \mathbf{x}_i'$$

从而，得到式(15.17)。

当回归元变化时，有：

$$\frac{\partial p_{ij}}{\partial \mathbf{x}_i} = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}_j}}{\sum_l e^{\mathbf{x}_i'\boldsymbol{\beta}_l}} \boldsymbol{\beta}_j - \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}_j}}{(\sum_l e^{\mathbf{x}_{il}'\boldsymbol{\beta}})^2} \sum_l e^{\mathbf{x}_i'\boldsymbol{\beta}_l} \boldsymbol{\beta}_l$$

$$= p_{ij}\boldsymbol{\beta}_j - p_{ij} \sum_l p_{il}\boldsymbol{\beta}_l = p_{ij}(\boldsymbol{\beta}_j - \bar{\boldsymbol{\beta}}_i)$$

其中，$\bar{\boldsymbol{\beta}}_l = \sum_l p_{il}\boldsymbol{\beta}_l$，正如式(15.19)所表述的。

15.12.3 嵌套 logit

考察由式(15.32)与式(15.33)给出的两水平 GEV 模型，满足：

$$G(\mathbf{Y}) = G(Y_{11}, \cdots, Y_{1K_1}, \cdots, Y_{J1}, \cdots, Y_{JK_J}) = \sum_{j=1}^{J} a_j \left(\sum_{k=1}^{K_j} Y_{jk}^{1/\rho_j}\right)^{\rho_j}$$

由于系数 a_j 的缘故，这是式(15.34)的推广。一般性 GEV 结果(15.31)变成 $\Pr[y_{jk}=1] = Y_{jk}G_{jk}/G(\mathbf{Y})$，其中，$G_{jk}$ 表示 $G(\mathbf{Y})$ 关于 Y_{jk} 的导数且在 $Y_{jk} = e^{V_{jk}}$ 处计算。

现在，有：

$$G_{jk} = \frac{\partial G(\mathbf{Y})}{\partial Y_{jk}} = a_j \left(\sum_{l=1}^{K_l} Y_{jl}^{1/\rho_j}\right)^{\rho_j - 1} \times Y_{jk}^{(1/\rho_j)-1}$$

从而，得出：

$$Y_{jk} G_{jk} = a_j \left(\sum_{l=1}^{K_l} Y_{jl}^{1/\rho_j}\right)^{\rho_j} \times Y_{jk}^{1/\rho_j}$$

于是：

$$p_{jk} \equiv \frac{Y_{jk}G_{jk}}{G(\mathbf{Y})} = \frac{a_j (\sum_{l=1}^{K_l} Y_{jl}^{1/\rho_j})^{\rho_j - 1} Y_{jk}^{1/\rho_j}}{\sum_{m=1}^{J} a_m (\sum_{l=1}^{K_l} Y_{ml}^{1/\rho_m})^{\rho_m}}$$

选取枝 j 的概率，在经过某些简化之后，得出：

$$p_j \equiv \sum_{k=1}^{K_j} p_{jk} = \frac{a_j (\sum_{l=1}^{K_l} Y_{jl}^{1/\rho_j})^{\rho_j}}{\sum_{m=1}^{J} a_m (\sum_{l=1}^{K_l} Y_{ml}^{1/\rho_m})^{\rho_m}}$$

而给定枝 j 选取分支 k 的条件概率是：

$$p_{k|j} \equiv \frac{p_{jk}}{p_j} = \frac{Y_{jk}^{1/\rho_j}}{\sum_{l=1}^{K_l} Y_{jl}^{1/\rho_j}}$$

此处结果，也是由马达拉(Maddala，1983，第 72 页)给出的。

我们需要在 $Y_{jk} = \exp(V_{jk})$ 处计算这些表达式。假定：

$$V_{jk} = \mathbf{z}_j'\boldsymbol{\alpha} + \mathbf{x}_{jk}'\boldsymbol{\beta}_j$$

于是，经过一些代数运算，得到：

$$(e^{V_{jk}})^{1/\rho_j} = \exp(\mathbf{z}_j'\boldsymbol{\alpha}/\rho_j)\exp(\mathbf{x}_{jk}'\boldsymbol{\beta}_j/\rho_j)$$

$$\sum_{l=1}^{K_l}(e^{V_{jl}})^{1/\rho_j} = \exp(\mathbf{z}_j'\boldsymbol{\alpha}/\rho_j)\exp(I_j)$$

$$\left(\sum_{l=1}^{K_l}(e^{V_{jl}})^{1/\rho_j}\right)^{\rho_j} = \exp(\mathbf{z}_j'\boldsymbol{\alpha}+\rho_j I_j)$$

其中：

$$I_j = \ln\left(\sum_{l=1}^{K_l}\exp(\mathbf{x}_{jk}'\boldsymbol{\beta}_j/\rho_j)\right)$$

由此可得，选取枝 j 的概率变成：

$$\begin{aligned}p_j &= \frac{a_j\left(\sum_{l=1}^{K_l}(e^{V_{jl}})^{1/\rho_j}\right)^{\rho_j}}{\sum_{m=1}^{J}a_m\left(\sum_{l=1}^{K_l}(e^{V_{ml}})^{1/\rho_m}\right)^{\rho_m}}\\ &= \frac{a_j\exp(\mathbf{z}_j'\boldsymbol{\alpha}+\rho_j I_j)}{\sum_{m=1}^{J}a_m(\exp(\mathbf{z}_m'\boldsymbol{\alpha}+\rho_m I_m))}\end{aligned}$$

这是如同式(15.36)第一项表述的。注意到，由于 $a_j\exp(\mathbf{z}_j'\boldsymbol{\alpha}+\rho_j I_j)=\exp(\ln a_j+\mathbf{z}_j'\boldsymbol{\alpha}+\rho_j I_j)$，所以纯量 a_j 能被并入 $\mathbf{z}_j$ 之中作为特定枝虚拟变量。因此，为了不失一般性，设 $a_j=1$。

位于枝 j 内分支 k 的概率是：

$$\begin{aligned}p_{k|j} &= \frac{(e^{V_{jk}})^{1/\rho_j}}{\sum_{l=1}^{K_l}(e^{V_{jl}})^{1/\rho_j}}\\ &= \frac{\exp(\mathbf{z}_j'\boldsymbol{\alpha}/\rho_j)\exp(\mathbf{x}_{jk}'\boldsymbol{\beta}/\rho_j)}{\sum_{l=1}^{K_l}\exp(\mathbf{z}_j'\boldsymbol{\alpha}/\rho_j)(\mathbf{x}_{jl}'\boldsymbol{\beta}/\rho_j)}\\ &= \frac{\exp(\mathbf{x}_{jk}'\boldsymbol{\beta}_j/\rho_j)}{\sum_{l=1}^{K_l}\exp(\mathbf{x}_{jl}'\boldsymbol{\beta}/\rho_j)}\end{aligned}$$

这是如同式(15.36)第二项表述的。

15.13 应用研究

多项式 logit 模型适用于描述数据或估计边缘概率，但因为独立于无关选项假设，如果需要参数的更多结构性解释，它被认为是一个不好的模型。许多软件包都有多项式 logit 模型估计。

运用 STATA 可估计嵌套 logit 模型，并运用依附于 LIMDEP 的 NLOGIT，而且很容易用诸如 GAUSS 语言编程。若存在明显的嵌套结构，就可使用这一模型，但通常不存在明显结构时。

随机参数 logit 模型要求用诸如 GAUSS 语言来特别编程，并需要运用第 12 章给出的基于模拟的估计方法。肯·特雷恩(Ken Train)在他的网站上提供了这方面的程序。

对于以上四种选择来说，估计多项式 probit 模型更会遇到挑战，而且相对而

言，在实证研究上获得成功的极少。鉴于上述原因，目前人们更偏爱随机参数 logit 模型。

15.14 文献注释

15.3 关于多项式模型的优秀参考书包括雨宫(Amemiya, 1981,1985)、马达拉(Maddala, 1983)以及格林(Greene, 2003)。本·阿基瓦和莱尔曼(Ben-Akiva and Lerman, 1985)、特雷恩(Train, 1986)以及伯尔施—祖潘(Borsch-Supan, 1987)都提供了广泛应用及理论综述。特雷恩(Train, 2003)对无序多项式模型与利用模拟方法的估计问题提供了优秀研究。

15.5 麦克法登(McFadden, 1981)的原创性文章，提供了离散选择建模的高等研究，并强调随机效用模型方法。对于福利分析，参见斯莫尔和罗森(Small and Rosen, 1981)、特雷恩(Train, 2003,第 59～61 页)以及达格斯文科和卡尔斯特罗姆(Dagsvik and Karström, 2004)。

15.6 伯尔施—祖潘(Borsch-Supan, 1987)对嵌套 logit 模型给出一个极好的解释及应用。

15.7 特雷恩(Train, 2003)的书还涵盖随机参数 logit 模型以及其他一些最新进展。雷维尔特和特雷恩(Revelt and Train, 1998)给出了一个早期应用。

15.8 博尔达克(Bolduc, 1999)给出了一个 9 种选择多项式 probit 模型的 MSL 估计。

习 题

15-1 考察由 $y^*=\mathbf{x}'\boldsymbol{\beta}+\varepsilon$ 建立的潜变量，其中 $\varepsilon\sim\mathcal{N}[0,1]$。假定当 $y^*<\alpha$ 时，观测到 $y=2$；当 $\alpha\leqslant y^*<U$ 时，观测到 $y=1$；同时当 $y^*\geqslant U$ 时，观测到 $y=0$，其中对每个个体而言，上限 U 是已知常数(即数据)，并且对于不同个体来说可能是不同的，不过 α 是未知的。

(a) 求 $y=0$、$y=1$ 以及 $y=2$ 的条件概率。

(b) 提供一致估计 $\boldsymbol{\beta}$ 与 α 方法的详细内容。

15-2 使用 15.2 节的钓鱼方式选择数据的 50%子样本。

(a) 估计 15.2.1 节的条件 logit 模型。

(b) 评论参数估计值的统计显著性。

(c) 各种钓鱼方式价格上涨的效应是多少?

15-3 使用 15.2 节的钓鱼方式选择数据的 50%子样本。

(a) 估计 15.2.2 节的多项式 logit 模型。

(b) 评论参数估计值的统计显著性。

(c) 各种钓鱼方式价格上涨的效应是多少?

15-4 使用 15.2 节的钓鱼方式选择数据的 50%子样本。假定我们将该模型合并成有三种选项的模型，同时对选项加以排序，若从码头或岸边钓鱼，则 $y=0$；

若从私家船钓鱼，则 $y=1$；而若租船钓鱼，则 $y=2$。

(a) 估计以收入作为唯一回归元的有序 logit 模型。

(b) 对估计系数给出解释。

(c) 把这个模型的拟合与以收入作为回归元的三种选择多项式模型的拟合加以比较。

16

Tobit 模型与选择模型

16.1 引 论

在本章，我们考察两个密切关联的专题：其一，关注的因变量是不完全观测的(**incompletely observed**)回归；其二，因变量是完全观测的，但观测上却处于并不代表总体的选择样本(**selected sample**)的回归。这包括受限因变量、潜变量、广义模型以及选择模型。

甚至在最简单的总体条件均值关于回归元为线性时，所有这些模型均享有共同的特征，OLS 回归导致非一致参数估计，原因在于样本不是总体的代表。一些可供选择的估计方法大部分均依赖于强分布假设，它们必须确保一致估计参数。

引起不完全观测数据的一些重要原因是：截尾与删失。对于截尾数据[1] (**truncated data**)来说，既有因变量中某些观测值的损失，又有回归元某些观测值的损失。例如，收入可能是因变量，而仅有低收入人员被包括在样本中。对于删失数据(**censored data**)来说，因变量信息会损失，但回归元数据却没有，例如，所有收入水平的人员都可能被包括在样本中，但为了保密，高收入人员的收入从上端进行编码，同时只报告大于它的信息，比如说每年 100 000 美元。与删失情况相比，截尾遭受更多信息损失。截尾与删失的一个重要例子是 Tobit 模型，它是以托宾(Tobin，1958)命名的，托宾在正态性下考察了线性回归。对于后面引进的其他模型的截尾与删失，会产生类似的问题，最著名的是第 17 章阐述的删失持续期限数据。更一般地，截尾与删失均是第 27 章将要研究的缺失数据问题的例子。

第一代估计方法需要强分布假设。当假定同方差误差时，甚至看似稍微违背假设，诸如异方差误差，都能导致非一致的参数估计。由于这种原因，本章所阐述的模型提供了半参数回归方法的经济计量学应用。对于删失与截尾的简单形式，比如上端编码来说，半参数方法得到了成功应用。不过，对于含有关于不可观测因素进行选择的更一般模型来说，到目前为止，还没有被广泛接受的方法。

16.2 节阐述删失与截尾的非线性回归模型的一般理论，而 16.3 节对 Tobit 模型进行专门研究。删失数据的一种可供选择模型是两部分模型，这在 16.4 节加以

[1] 又称为截断数据。——译者注

阐述。16.5 节阐述样本选择模型。16.6 节关于健康消费支出的应用讨论，与两部分模型及样本选择模型形成了对比。不可观测的反事实框架的罗伊模型将在 16.7 节阐述。16.8 节考察完全结构模型，这可通过含有角点解的效用最大化，或通过把联立方程模型扩展到选择样本上而获得。

16.2 删失模型与截尾模型

我们阐述，当数据是删失或截尾的时候，对完全参数模型进行估计的一般方法。这些方法能用于后面几章将要阐述的一些模型，例如计数模型与持续期限模型。重要的例子是线性模型中关于删失或截尾的模型，这在 16.2 节引进，并在 16.3 节给出各自研究。

16.2.1 *删失与截尾例子*

设 y^* 表示不完全可观测的变量。对于从下面的截尾来说，当 y^* 大于某个门限值时，y^* 才是可观测的。为了简单起见，设那个门限值为 0。于是，当 $y^*>0$ 时，我们观测到 $y=y^*$。由于负值没有出现在样本中，故截尾均值大于 y^* 的均值。对于从下面 0 点处的删失来说，当 $y^*\leqslant 0$ 时，y^* 不是完全可观测的，却知道 $y^*<0$，于是，为了简单起见，令 y 等于 0。由于负值标度至多为零，所以删失均值同样会大于 y^* 的均值。很明显，为了估计最初总体均值，截尾与删失样本的样本均值若没有调整，就不能加以运用。

本章将研究回归模型的类似问题。幸运的是，一旦令斜率系数不变，截尾与删失可能只会导致截距的上下移动；不过，情况还远不止这些。例如，若最初模型为 $\mathrm{E}[y^*|\mathbf{x}]=\mathbf{x}'\boldsymbol{\beta}$，则截尾或删失导致关于 $\mathbf{x}$ 与 $\boldsymbol{\beta}$ 为非线性的，所以 OLS 给出了 $\boldsymbol{\beta}$ 的非一致估计，从而产生边际效应的非一致估计。

举一个例子阐述，考察下面利用模拟数据的劳动力供给例子。对人们希望的年度工作小时数 y^* 与计时工资 w 之间的关系，设定成具有线性对数的形式，满足数据生成过程：

$$\begin{aligned} y^* &= -2\,500+1\,000\ln w+\varepsilon \\ \varepsilon &\sim \mathcal{N}[0,\ 1\,000^2] \\ \ln w &\sim \mathcal{N}[2.75,\ 0.60^2] \end{aligned} \tag{16.1}$$

这是一个 Tobit 模型，对它的详细研究在 16.3 节给出。该模型蕴含着工资弹性是 $1\,000/y^*$，例如，这等于全日制工作(2 000 小时)。工资每增加 10%，年度工作小时增加 10 个小时。

图 16.1 显示关于 200 个观测值生成样本的 $\ln w$ 与 y^* 的散点图，关于 y^* 的无条件均值为 $-2\,500+1\,000\ln w$，该值由最下面曲线给出，它是一条直线。

对于在 0 点删失来说，将 y^* 的负值设为 0，因为具有负的工作意愿小时的人员不会去工作。对于这种特殊样本来说，它约为观测值的 35%。这促使低工资均值上移，因为 y^* 的许多负值被移至 0。它对高工资很少有影响，从那时起 y^* 上很少

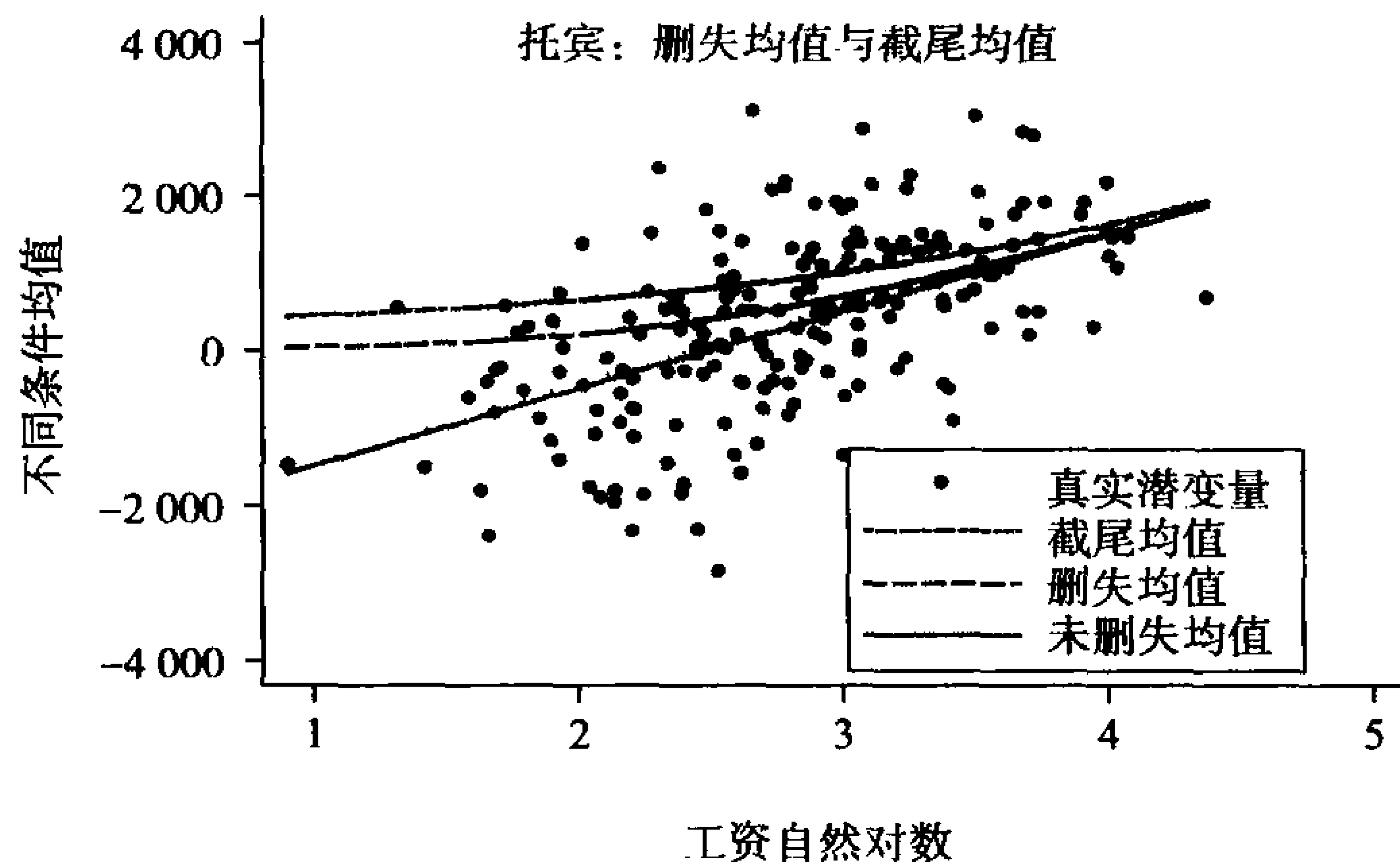


图 16.1 小时对工资对数的回归：未删失均值（下面）、删失条件均值（中间），以及关于在 0 点小时处以下删失/截尾的截尾条件均值（上面）。数据由经典线性回归模型生成。

观测值为 0。利用后面的式(16.23)，图 16.1 的中间曲线给出了作为结果的截尾均值。

很明显，即使基本总体均值是线性的，删失与截尾的条件均值均关于 **x** 为非线性的。利用截尾或删失数据的 OLS 估计将导致斜率参数的非一致估计，观察可知，图 16.1 与原来未截尾的均值相比，对非线性截尾与删失的线性近似将具有较平坦的斜率。相反，分析应建立在截尾或删失的条件均值的基础上。不幸的是，正如我们将要看到的，这些均建立在强分布的假设基础上。

对于在 0 点截尾来说，y^* 含有负值总体的 35%都被省略。这使得其均值大于删失均值，因为 0 值不再被包括在数据之中用于构成均值。当利用后面的式(16.23)，图 16.1 上的曲线给出了作为结果的截尾均值。

16.2.2 *删失与截尾机制*

作为回归分析的一种习惯，设 y 表示因变量的观测值。违背通常分析的是，y 成为潜变量(**latent dependent variable**)y^* 的不完全观测值，其中，对于某个设定函数 $g(\cdot)$来说，观测规则是：

$$y=g(y^*)$$

$g(\cdot)$ 的一些重要例子如下。

删失

对于删失，我们总是观测到回归元 **x**，就 y^* 可能值的子集而言，完全观测到 y^*，而就 y^* 的其余可能值而言，不完全观测到 y。当删失从下面(或从左侧)进行，会观测到：

$$y=\begin{cases}y^*, & 当\ y^*>L \\ L, & 当\ y^*\leqslant L\end{cases} \tag{16.2}$$

例如，对于某些耐用品支出为正值($y^*>0$)的人与其他拥有零支出($y^*\leqslant 0$)的人，

所有消费者均可能被抽样。当删失从上面(或从右侧)进行,会观测到:

$$y=\begin{cases}y^*, & \text{当 } y^*<U \\ U, & \text{当 } y^*\geqslant U\end{cases} \tag{16.3}$$

例如,年收入数据可能是在 $U=100\,000$ 美元处进行上端编码。删失的这种形式在持续期限文献中称为第 I 类删失(参见 17.4.1 节)。

为了简单起见,将 y^* 的不完全可观测的观测值设为 L 或 U。更一般地,我们要求知道,不完全可观测的观测值 y^* 是缺失的(也就是说,可观测到 y^* 位于有关界限之外),并且回归元 **x** 继续是完全可观测的。

截尾

由于所有观测值数据在某个界限处丢失,所以截尾承受额外的信息损失。对于从下面截尾来说,仅仅观测到:

$$\text{当 } y^*>L \text{ 时},\ y=y^* \tag{16.4}$$

例如,只对购买耐用品的消费者进行抽样。对于从上面截尾来说,我们仅仅观测到:

$$\text{当 } y^*<U \text{ 时},\ y=y^* \tag{16.5}$$

例如,只对低收入个体进行抽样。

区间数据

区间数据(**interval data**)是以区间形式记录的数据。调查数据经常是以这种方式收集的,以便帮助回忆并提供某个较大匿名者答复更多的个人问题。例如,收入可能被报告在 10 000 美元至上端编码 100 000 美元处。这类数据在多个点处被删失,观测到数据 y 位于某个特殊区间之内,不可观测值 y^* 就位于该区间内。

16.2.3 *删失与截尾*

倘若研究者应用完全参数方法,则对删失与截尾很容易处理。例如,这可能是含有区间数据或上端编码数据的情形,合理假定收入为对数正态分布或医生出诊次数为负的二项分布。

如果对给定回归元时 y^* 的条件分布加以设定,那么这种分布的参数能通过基于删失或截尾 y 的条件分布的 ML 估计而得到一致且有效的估计。特别地,设 $f^*(y^*|\mathbf{x})$ 与 $F^*(y^*|\mathbf{x})$ 表示潜变量 y^* 的条件概率密度函数(或者概率质量函数)与累积分布函数。于是,由于 $y=g(y^*)$ 是 y^* 的变换,所以人们总是能获得其可观测因变量 y 的对应条件 pdf 与 cdf。

参数方法的局限性是它依据强分布假设。例如,对于线性回归模型来说,即使误差是非正态的,在正态性下 MLE 仍是保持一致的,但若误差是非正态的,删失会变成非一致的(参见 16.3.2 节)。更为灵活的一些模型与半参数方法将在后面几节阐述。

删失 MLE

删失与截尾既可使条件均值变化,又可使条件密度变化。我们下面以密度开

始研究。

考察给定从下面删失的 ML 估计。对于 $y>L$，y 的密度与 y^* 的密度是一样的，所以 $f(y|\mathbf{x})=f^*(y|\mathbf{x})$。对于 $y=L$，即下界情况，其密度是含有质量等于观测 $y^*\leqslant L$ 概率的离散或 $F^*(L|\mathbf{x})$。因而，对于下面删失来说：

$$f(y|\mathbf{x})=\begin{cases}f^*(y|\mathbf{x}), & 当\ y>L\\ F^*(L|\mathbf{x}), & 当\ y=L\end{cases}$$

正如式(16.3)所提及的，当 $y^*\leqslant L$ 时，不一定设 $y=L$。当 $y^*\leqslant L$ 时，即使没有可观测的 y 值，其密度仍然是 $F^*(L|\mathbf{x})$。

密度是 y^* 的 pdf 与 cdf 的混合之物。与二值结果模型分析相似，在记号形式上引入指示变量

$$d=\begin{cases}1, & 当\ y>L\\ 0, & 当\ y=L\end{cases} \tag{16.6}$$

是方便的。于是，给定从下面删失时，条件密度能重新写成：

$$f(y|\mathbf{x})=f^*(y|\mathbf{x})^d F^*(L|\mathbf{x})^{1-d} \tag{16.7}$$

对于 N 个独立观测值的样本，其删失 MLE 是对

$$\ln \mathrm{L}_N(\boldsymbol{\theta}) = \sum_{i=1}^{N}\{d_i \ln f^*(y_i|\mathbf{x}_i,\boldsymbol{\theta}) + (1-d_i)\ln F^*(L_i|\mathbf{x}_i,\boldsymbol{\theta})\} \tag{16.8}$$

求极大值，其中，$\boldsymbol{\theta}$ 表示 y^* 分布的参数。一般来讲，删失下界 L_i 被允许随不同个体而变化，尽管通常 $L_i=L$。倘若未删失变量的最初密度 $f^*(y^*|\mathbf{x},\boldsymbol{\theta})$被正确设定，则删失 MLE 是一致的且渐近正态的。

然而，当删失是从上面进行时，对数似然类似于式(16.8)，现在只是当 $y<U$ 时 $d=1$，否则 $d=0$，同时用 $1-F^*(U|\mathbf{x},\boldsymbol{\theta})$ 代替 $F^*(L|\mathbf{x},\boldsymbol{\theta})$。一个重要例子是，右删失持续期限数据(参见 17.4 节)。

截尾 MLE

对于在 L 处从下面截尾来说，不使用对 $\mathbf{x}$ 的相依性，观测 y 的条件密度为：

$$\begin{aligned}f(y)&=f^*(y|y>L)\\&=f^*(y)/\Pr[y|y>L]\\&=f^*(y)/[1-F^*(L)]\end{aligned}$$

因此，其截尾 MLE 是对

$$\ln \mathrm{L}_N(\boldsymbol{\theta}) = \sum_{i=1}^{N}\{\ln f^*(y_i|\mathbf{x}_i,\boldsymbol{\theta}) - \ln[1-F^*(L_i|\mathbf{x}_i,\boldsymbol{\theta})]\} \tag{16.9}$$

求极大值。相反，如果截尾是从上面进行的，那么对数似然是式(16.9)，只是用 $F^*(U|\mathbf{x},\boldsymbol{\theta})$代替 $1-F^*(L|\mathbf{x},\boldsymbol{\theta})$。

若忽略删失或截尾，则会导致非一致性。例如，若截尾被忽略了，MLE 对 $\sum_i \ln f^*(y_i|\mathbf{x}_i,\boldsymbol{\theta})$求极大值，这样做得到错误的似然函数，因为它省略式(16.9)中

的第二项。删失与截尾的一致性要求对 $f(\cdot)$ 正确设定，从而要求对 $f^*(\cdot)$潜变量密度正确设定。即使 $f^*(\cdot)$是 LEF 密度(参见 5.7.3 节)，若阐述删失或截尾，则不仅均值必须得到正确设定，其密度也必须要得到正确设定。

区间数据 MLE

假定潜变量只有位于互不相交区间$(-\infty,a_1]$，$(a_1,a_2]$，…，(a_J,∞)时，才是可观测的，其中 $a_1,a_2,\cdots,a_J$ 均是已知的。于是，由于：

$$\begin{aligned}\Pr[a_j<y^*\leqslant a_{j+1}]&=\Pr[y^*\leqslant a_{j+1}]-\Pr[y^*\leqslant a_j]\\&=F^*(a_{j+1})-F^*(a_j)\end{aligned}$$

所以区间数据 MLE 是对

$$\ln \mathrm{L}_N(\boldsymbol{\theta})=\sum_{i=1}^{N}\sum_{j=0}^{J}d_{ij}\ln[F^*(a_{j+1}\mid\mathbf{x}_i,\boldsymbol{\theta})-F^*(a_j\mid\mathbf{x}_i,\boldsymbol{\theta})] \qquad (16.10)$$

求极大值，其中，d_{ij}表示二值指示变量，$j=0,\cdots,J$，当 $y_{ij}\in(a_j,a_{j+1}]$时，$d_{ij}=1$，否则为 0。这类似于有序 probit 或者 logit 模型(参见 15.9.1 节)，只是此处区间边界 $a_1,\cdots,a_J$ 均是已知的。

16.2.4 *泊松删失与截尾 MLE 例子*

假定服从泊松分布，因此 $f^*(y)=e^{-\mu}\mu^y/y!$，并且 $\ln f^*(y)=-\mu+y\ln\mu-\ln y!$，其均值 $\mu=\exp(\mathbf{x}'\boldsymbol{\beta})$。

假定对去健康诊所就诊的次数进行建模，但只有那些去健康诊所就诊的人员数据才可以利用。于是，此数据是从 0 处以下截尾的，同时当 $y^*>0$ 时，我们才可观测到 $y=y^*$。从而，$F^*(0)=\Pr[y^*\leqslant 0]=\Pr[y^*=0]=e^{-\mu}$，由式(16.9)知，关于 $\boldsymbol{\beta}$ 的截尾 MLE 是对

$$\ln \mathrm{L}_N(\boldsymbol{\beta})=\sum_{i=1}^{N}\{-\exp(\mathbf{x}_i'\boldsymbol{\beta})+y_i\mathbf{x}_i'\boldsymbol{\beta}-\ln y_i!-\ln[1-\exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta}))]\}$$

求极大值。

然而，假定由于上端编码的缘故，数据在从上面 10 处被删失，因此，当 $y^*<10$ 时，我们观测到 $y=y^*$，而当 $y^*\geqslant 10$ 时，$y=10$，从而，$\Pr[y^*\geqslant 10]=1-\Pr[y^*<10]=1-\sum_{k=0}^{9}f^*(k)$。由式(16.8)知，关于 $\boldsymbol{\beta}$的删失 MLE 是对

$$\begin{aligned}\ln \mathrm{L}_N(\boldsymbol{\beta})=\sum_{i=1}^{N}\Big\{&d_i[-\exp(\mathbf{x}_i'\boldsymbol{\beta})+y_i\mathbf{x}_i'\boldsymbol{\beta}-\ln y_i!]\\&+(1-d_i)\ln\Big[\sum_{k=0}^{9}e^{-\exp(\mathbf{x}_i'\boldsymbol{\beta})}(\exp(\mathbf{x}_i'\boldsymbol{\beta}))^k/k!\Big]\Big\}\end{aligned}$$

求极大值。

在上述两种情况下，与没有截尾或删失的那些泊松一阶条件相对比，得到的一阶条件是相当复杂的。再者，在这两种情况下，若忽略截尾或删失，并求最初密度极大值，则导致非一致参数估计。

16.2.5 删失与截尾条件均值

删失与截尾都促使条件均值改变了。

例如，考察从下面 0 处截尾的泊松分布。其截尾密度为 $f^*(y)/[1-F^*(0)]$，$y=1,2,\cdots$，所以截尾均值为 $\sum_{k=1}^{\infty}kf^*(k)/[1-F^*(0)]=\sum_{k=0}^{\infty}kf^*(k)/[1-F^*(0)]=\mu/(1-e^{-\mu})$。因而有：

$$\mathrm{E}[y|\mathbf{x}]=\exp(\mathbf{x}'\boldsymbol{\beta})/[1-\exp(-\exp(\mathbf{x}'\boldsymbol{\beta}))]$$

而不是没有截尾情形的 $\exp(\mathbf{x}'\boldsymbol{\beta})$。

$\mathrm{E}[y|\mathbf{x}]$的这个表达式可用于 NLS 估计。不过，由于给定截尾时 NLS 估计量依赖的分布假设在本质上强于更有效 ML 估计量一致性需要的那些假设，所以相对于 NLS 而不是 ML 估计来说，并没有什么优势。

16.3 Tobit 模型

在包括服从正态分布误差的经济计量学线性回归模型中，当只有正的结果是完全可观测的时候，最常出现截尾与删失。这种模型以托宾（Tobin, 1958）名字命名，托宾将它应用于消费者耐用品的个体开支上。实际上，该模型通常表现出约束性太强。不过，这里仍以某种详细方式阐述它，因为此模型为本章后面几节要阐述的更一般模型提供了基础。

16.3.1 *Tobit* 模型

删失正态回归模型或 Tobit 模型，是一种从下面 0 点处删失的模型，其中，潜变量关于回归元是线性的，其可加误差是正态分布的且同方差的。因而，有：

$$y^*=\mathbf{x}'\boldsymbol{\beta}+\varepsilon \tag{16.11}$$

其中，误差项为：

$$\varepsilon\sim\mathcal{N}[0,\sigma^2] \tag{16.12}$$

对不同观测值来说，具有常值方差 σ^2。这蕴含，潜变量 $y^*\sim\mathcal{N}[\mathbf{x}'\boldsymbol{\beta},\sigma^2]$。可观测的 y 是满足 $L=0$ 的。这一结果由式(16.2)定义，因此，有：

$$y=\begin{cases}y^*, & \text{当 } y^*>0\\ -, & \text{当 } y^*\leqslant 0\end{cases} \tag{16.13}$$

其中，“－”表示 y 作为缺失可观测的。当 $y^*\leqslant 0$ 时，y 的特殊值没有必要一定是可观测的，尽管在某些设置下，例如耐用品支出，我们可以观测到 $y=0$。

方程组(16.11)～(16.13)定义出由托宾（Tobin, 1958）分析的原始 Tobit 模型。更一般地，Tobit 模型是以式(16.11)与式(16.12)关于潜变量开始的，但可拥有其他的删失机制，包括从上面删失、从下面与上面删失（两部分限制 Tobit 模型），以及区间删失数据。本节中的结果被限制在由式(16.13)给出的删失机制上。

后面几节模型，有时称为广义 Tobit 模型。

在许多设置背景下，正规化 $L=0$ 不仅是自然的，而且对含有截距且常值门限参数的线性模型来说是必需的。于是，当 $y^*>L$ 时，或者等价地当 $\beta_1+\mathbf{x}_2'\boldsymbol{\beta}_2+\varepsilon>L$ 或 $(\beta_1-L)+\mathbf{x}_2'\boldsymbol{\beta}_2+\varepsilon>0$ 时，才可观测到 y。因而，唯一的差 (β_1-L) 是可识别的。更一般地，从观测上看，含有变量删失门限 $L=\mathbf{x}'\boldsymbol{\gamma}$ 的潜模型 $y^*=\mathbf{x}'\boldsymbol{\beta}+\varepsilon$ 等价于含有固定门限 $L=0$ 的潜模型 $y^*=\mathbf{x}'(\boldsymbol{\beta}-\boldsymbol{\gamma})+\varepsilon$。这些结果是含有可加误差线性模型因删失产生的重要结论，却不可应用到非线性模型上，譬如前面的泊松例子。

若对删失密度应用一般表达式(16.7)，则此处 $f^*(y)$ 是 $\mathcal{N}[\mathbf{x}'\boldsymbol{\beta},\sigma^2]$ 密度，同时有：

$$\begin{aligned}F^*(0)&=\Pr[y^*\leqslant 0]\\&=\Pr[\mathbf{x}'\boldsymbol{\beta}+\varepsilon\leqslant 0]\\&=\Phi(-\mathbf{x}'\boldsymbol{\beta}/\sigma)\\&=1-\Phi(\mathbf{x}'\boldsymbol{\beta}/\sigma)\end{aligned}$$

其中，$\Phi(\cdot)$ 表示标准正态 cdf，而最后等式使用了标准正态分布的对称性。因而，删失密度能表述成：

$$f(y)=\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(y-\mathbf{x}'\boldsymbol{\beta})^2\right\}\right]^d\left[1-\Phi\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right)\right]^{1-d}\qquad(16.14)$$

其中，二值指示变量是由式(16.6)定义的，并满足 $L=0$。

Tobit MLE $\hat{\boldsymbol{\theta}}=(\hat{\boldsymbol{\beta}}',\hat{\sigma}^2)'$ 是对删失对数似然函数(16.8)求极大值。给定式(16.14)，它变成：

$$\begin{aligned}\ln \mathrm{L}_N(\boldsymbol{\beta},\sigma^2)=\sum_{i=1}^{N}\Big\{&d_i\left(-\frac{1}{2}\ln 2\pi-\frac{1}{2}\ln\sigma^2-\frac{1}{2\sigma^2}(y_i-\mathbf{x}_i'\boldsymbol{\beta})^2\right)\\&+(1-d_i)\ln\left(1-\Phi\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right)\Big\}\end{aligned}\qquad(16.15)$$

它为离散密度与连续密度的混合。其一阶条件为：

$$\begin{aligned}\frac{\partial\ln \mathrm{L}_N}{\partial\boldsymbol{\beta}}&=\sum_{i=1}^{N}\frac{1}{\sigma^2}\left(d_i(y_i-\mathbf{x}_i'\boldsymbol{\beta})-(1-d_i)\frac{\sigma\phi_i}{(1-\Phi_i)}\right)\mathbf{x}_i=\mathbf{0}\\\frac{\partial\ln \mathrm{L}_N}{\partial\sigma^2}&=\sum_{i=1}^{N}\left\{d_i\left(-\frac{1}{2\sigma^2}+\frac{(y_i-\mathbf{x}_i'\boldsymbol{\beta})^2}{2\sigma^4}\right)+(1-d_i)\frac{\phi_i\mathbf{x}_i'\boldsymbol{\beta}}{(1-\Phi_i)}\frac{1}{2\sigma^3}\right\}=0\end{aligned}\qquad(16.16)$$

这里利用了 $\partial\Phi(z)/\partial(z)=\phi(z)$，其中，$\phi(\cdot)$ 表示标准正态 pdf，而且满足定义 $\phi_i=\phi(\mathbf{x}_i'\boldsymbol{\beta}/\sigma)$ 且 $\Phi_i=\Phi(\mathbf{x}_i'\boldsymbol{\beta}/\sigma)$。与以往一样，如果密度得到正确设定，即数据生成过程是式(16.11)与式(16.12)，且删失机制是式(16.13)，那么 $\hat{\boldsymbol{\theta}}$ 是一致的。MLE 服从渐近正态分布，例如，其方差矩已由马达拉(Maddala，1983，第 155 页)和雨宫(Amemiya，1985，第 373 页)给出。

托宾(Tobin，1958)提出 Tobit 模型的 ML 估计，同时断言可应用 ML 理论。雨宫(Amemiya，1973)提供了通常理论可以应用的正式证明，尽管删失密度具有混合的离散—连续特性。雨宫在这篇经典论文附录中详述了由 5.3 节阐述的极值

估计量。

如果数据是从下面 0 点处截尾的而不是删失的，那么 Tobit MLE 是对截尾正态对数似然函数：

$$\ln \mathrm{L}_N(\boldsymbol{\beta},\sigma^2)=\sum_{i=1}^{N}\left\{-\frac{1}{2}\ln\sigma^2-\frac{1}{2}\ln 2\pi-\frac{1}{2\sigma^2}(y_i-\mathbf{x}_i'\boldsymbol{\beta})^2-\ln\Phi(\mathbf{x}_i'\boldsymbol{\beta}/\sigma)\right\} \tag{16.17}$$

求极大值，一旦利用关于 y^* 的式(16.9)，可获得如同式(16.11)与式(16.12)的分布。

16.3.2 *Tobit MLE* 的非一致性

Tobit MLE 的一个非常严重的弱点，是它紧密地依赖于分布假设。若误差 ε 是异方差的或非正态的，MLE 则为非一致的。

这可从 MLE 一阶条件(16.16)看出，它是包含变量 d_i、y_i、ϕ_i 和 Φ_i 的复杂函数。式(16.16)中第一个方程满足 $\mathrm{E}[\partial\ln \mathrm{L}_N/\partial\boldsymbol{\beta}]=\mathbf{0}$，即一致性的必要条件（参见 5.3.7 节），倘若：

$$\mathrm{E}[d_i]=\Phi_i$$
$$\mathrm{E}[d_i\, y_i]=\Phi_i\mathbf{x}_i'\boldsymbol{\beta}+\sigma\phi_i$$

可以证明，若数据生成过程是式(16.11)与式(16.12)，且删失机制为式(16.13)，则这些矩条件成立。不过，在数据生成过程的任何其他设定下，它们不可能成立，因为其紧密地依赖于正态性和同方差性。例如，具有异方差误差，此估计量是非一致的，从而 $\mathrm{E}[d_i]=\Phi(\mathbf{x}_i'\boldsymbol{\beta}/\sigma_i)\neq\Phi_i$，除非 $\sigma_i^2=\sigma^2$。

通过对异方差性设定一个模型，比如说 $\sigma_i^2=\exp(\mathbf{z}_i'\boldsymbol{\gamma})$，对带有异方差正态误差的模型进行一致估计是可能的。对于从 0 点处下面删失，对数似然 $\ln \mathrm{L}_N(\boldsymbol{\beta},\boldsymbol{\gamma})$ 是由式(16.15)给出的，并且用 $\exp(\mathbf{z}_i'\boldsymbol{\gamma})$ 代替 σ^2。于是，一致性需要正态误差，且对异方差性的函数形式要求正确设定。

很明显，就删失或截尾而言，分布假设变得极为重要，甚至在没有删失或没有截尾的情况下，对错误设定稍微稳健的分布也是如此。对 Tobit 模型的设定检验在 16.3.7 节加以讨论。在许多删失数据应用中，Tobit 模型并不合适。相反，要运用本章后面几节阐述的更一般模型。

16.3.3 线性回归的删失与截尾均值

线性回归模型(16.11)中的删失与截尾，会导致可观测因变量 y 拥有含有条件均值而不是 $\mathbf{x}'\boldsymbol{\beta}$ 的分布，即使 ε 是同方差的，其方差为条件方差而不是 σ^2，同时即使 ε 是正态分布，该分布也是非正态的。

在专门研究 16.3.4 节至 16.3.7 节的正态分布误差之前，我们在本节阐述线性回归的一般结果。一些结果提供了关于截尾与删失后果的另一种见解，同时形成后面几节阐述的非估计方法的基础。

我们以截尾均值开始。从直观上看，截尾的影响是可预测的。在截尾排除很

小的值，因此其均值应增大，而右截尾的均值应减小。由于截尾使变异范围缩小，所以其方差应减小。

对于在 0 点左截尾来说，当 $y^*>0$ 时，才观测到 y。为了记号简单起见，我们不用对 x 期望的相依性，那么在截尾均值变成：

$$\begin{aligned}\mathrm{E}[y]&=\mathrm{E}[y^*|y^*>0]\\&=\mathrm{E}[\mathbf{x}'\boldsymbol{\beta}+\varepsilon|\mathbf{x}'\boldsymbol{\beta}+\varepsilon>0]\\&=\mathrm{E}[\mathbf{x}'\boldsymbol{\beta}|\mathbf{x}'\boldsymbol{\beta}+\varepsilon>0]+\mathrm{E}[\varepsilon|\mathbf{x}'\boldsymbol{\beta}+\varepsilon>0]\\&=\mathbf{x}'\boldsymbol{\beta}+\mathrm{E}[\varepsilon|\varepsilon>-\mathbf{x}'\boldsymbol{\beta}]\end{aligned}\tag{16.18}$$

其中，第二个等式使用了式(16.11)，并且最后等式假定 ε 与 $\mathbf{x}$ 是独立的。如同人们所料，截尾均值大于 $\mathbf{x}'\boldsymbol{\beta}$，因为对任何常值 c 来说，$\mathrm{E}[\varepsilon|\varepsilon>c]$将大于 $\mathrm{E}[\varepsilon]$。

对于在 0 点的左删失来说，假定可观测到 $y=0$，而不仅是 $y^*\leqslant 0$。删失均值可通过首先对可观测的 y 以满足 $L=0$ 并由式(16.6)所定义的二值指示变量为条件，然后无条件化。为了记号简单起见，再一次不用对 $\mathbf{x}$ 的相依性，我们得到左删失均值：

$$\begin{aligned}\mathrm{E}[y]&=\mathrm{E}_d[\mathrm{E}_{y|d}[y|d]]\\&=\Pr[d=0]\times\mathrm{E}[y|d=0]+\Pr[d=1]\times\mathrm{E}[y|d=1]\\&=0\times\Pr[y^*\leqslant 0]+\Pr[y^*>0]\times\mathrm{E}[y^*|y^*>0]\\&=\Pr[y^*>0]\times\mathrm{E}[y^*|y^*>0]\end{aligned}\tag{16.19}$$

其中，$\Pr[y^*>0]=1-\Pr[y^*\leqslant 0]=\Pr[\varepsilon>-\mathbf{x}'\boldsymbol{\beta}]$表示 1 减去删失概率，而 $\mathrm{E}[y^*|y^*>0]$表示由式(16.18)推导的截尾均值。

总之，对于线性回归模型来说，从下面 0 点处删失或截尾，其条件均值由

$$\begin{aligned}&\text{潜变量：} && \mathrm{E}[y^*|\mathbf{x}]=\mathbf{x}'\boldsymbol{\beta}\\&\text{左截尾(在 0 点)：} && \mathrm{E}[y|\mathbf{x},y>0]=\mathbf{x}'\boldsymbol{\beta}+\mathrm{E}[\varepsilon|\varepsilon>-\mathbf{x}'\boldsymbol{\beta}]\\&\text{左删失(在 0 点)：} && \mathrm{E}[y|\mathbf{x}]=\Pr[\varepsilon>-\mathbf{x}'\boldsymbol{\beta}]\{\mathbf{x}'\boldsymbol{\beta}+\mathrm{E}[\varepsilon|\varepsilon>-\mathbf{x}'\boldsymbol{\beta}]\}\end{aligned}\tag{16.20}$$

给出。很显然，尽管最初条件均值是线性的，但删失或截尾会导致条件均值是非线性的，因此 OLS 估计将是非一致的。

所采用的一种可能方法是对 ε 的分布假定成参数形式。这会产生 $\mathrm{E}[\varepsilon|\varepsilon>-\mathbf{x}'\boldsymbol{\beta}]$与 $\Pr[\varepsilon>-\mathbf{x}'\boldsymbol{\beta}]$ 的表达式，从而获得截尾或删失条件均值。我们在下一节关于正态分布误差的条件下解决此问题。

第二个方法试图回避或极小化这类参数假设。我们在下一节将讨论该问题，但这里注意到，不管 ε 的分布如何，由于 $\mathrm{E}[\varepsilon|\varepsilon>-\mathbf{x}'\boldsymbol{\beta}]$关于 $\mathbf{x}'\boldsymbol{\beta}$是单调递减函数，所以截尾均值是含有关于 $\mathbf{x}'\boldsymbol{\beta}$是递减的校正项的单指标模型。

16.3.4 *Tobit 模型的删失与截尾均值*

对于 Tobit 模型来说，回归误差 ε 是正态的，并且我们运用将由 16.10.1 节推导的下述结果。

命题 16.1(标准正态的截尾矩) *假定 $z\sim\mathcal{N}[0,1]$。于是，z 的左截尾矩是：*

(i) $E[z|z>c]=\phi(c)/[1-\Phi(c)]$ 并且 $E[z|z>-c]=\phi(c)/\Phi(c)$；

(ii) $E[z^2|z>c]=1+c\phi(c)/[1-\Phi(c)]$；

(iii) $V[z|z>c]=1+c\phi(c)/[1-\Phi(c)]-\phi(c)^2/[1-\Phi(c)]^2$。

命题 16.1 的结果(i)已表示在图 16.2 中。我们考察 $z\sim\mathcal{N}[0,1]$ 从下面 0 点截尾，其中，c 可从 -2 到 2 变动。最下面的曲线是在 c 点计算的标准正态密度。中间曲线是在 c 点计算的标准正态 cdf $\Phi(c)$，同时当在 c 点截尾时，给出了截尾的概率。这一概率在 $c=-2$ 时大致为 0.023，而在 $c=2$ 时大致为 0.977。最上面的曲线给出截尾均值 $E[z|z>c]=\phi(c)/[1-\Phi(c)]$。如同所预期的，对于 $c=-2$，这接近于 $E[z]=0$，从而几乎没有截尾，而且 $E[z|z>c]>c$。没有预料到的是，先验为 $\phi(c)/[1-\Phi(c)]$，特别对 $c>0$，$\phi(c)/[1-\Phi(c)]$ 大致是线性的。当截尾从上面进行时，就可利用矩，例如，$E[z|z<c]=-E[-z|-z>-c]=-\phi(c)/\Phi(c)$。

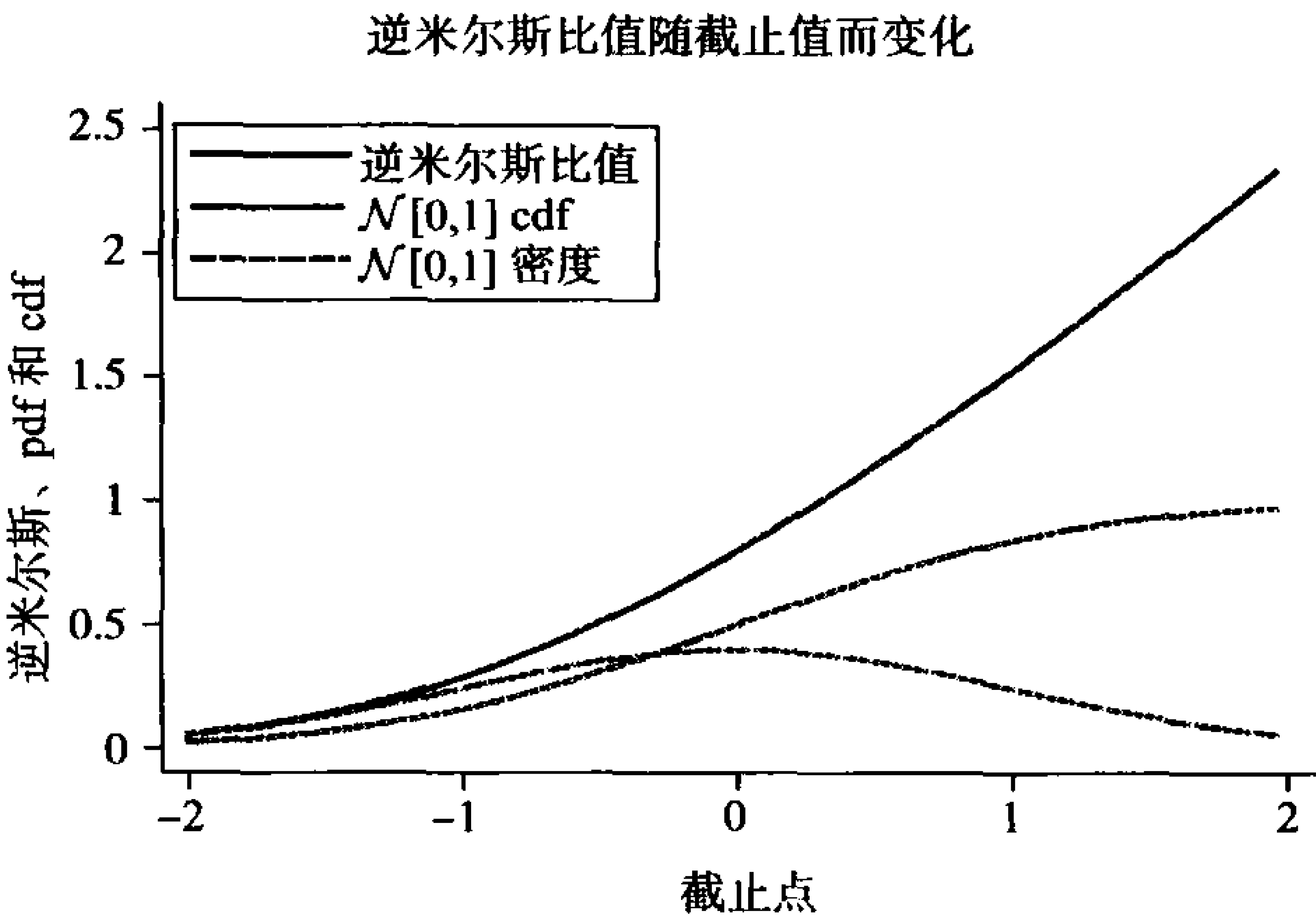


图 16.2 当删失或截止点 c 增大时，标准正态分布的逆为米尔斯比值。同时画出标准正态 cdf 与密度。

若把这一结果应用于式(16.18)，则误差项具有删失均值：

$$\begin{aligned}E[\varepsilon|\varepsilon>-\mathbf{x}'\boldsymbol{\beta}]&=\sigma E\left[\frac{\varepsilon}{\sigma}\middle|\frac{\varepsilon}{\sigma}>\frac{-\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right]\\&=\sigma\phi\left(-\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right)\Big/\left[1-\Phi\left(-\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right)\right]\\&=\sigma\phi\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right)\Big/\left[\Phi\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right)\right]\\&=\sigma\lambda\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right)\end{aligned}\tag{16.21}$$

其中，第二个等式使用了命题 16.1，第三个行使用了 $\phi(z)$ 关于 0 的对称性，而且我们定义：

$$\lambda(z)=\frac{\varphi(z)}{\Phi(z)}\tag{16.22}$$

如同式(16.22)在定义时遵循了雨宫(Amemiya，1985)以及许多其他学者的定义及术语，将它称为逆米尔斯比值(**inverse Mills ratio**)。由约翰逊和卡茨(Johnson

and Kotz, 1970, 第 278 页)可知,米尔斯实际上将比值$(1-\Phi(z))/\phi(z)$列成表,而该比值的逆是正态分布的风险函数。因此,一些作者反而把式(16.21)写成$\mathrm{E}[\varepsilon|\varepsilon>-\mathbf{x}'\boldsymbol{\beta}]=\sigma\lambda^*(-\mathbf{x}'\boldsymbol{\beta}/\sigma)$,其中,将$\lambda^*(z)=\phi(z)/\Phi(-z)$称为逆米尔斯比值。

同理,$\Pr[\varepsilon>-\mathbf{x}'\boldsymbol{\beta}]=\Pr[-\varepsilon<\mathbf{x}'\boldsymbol{\beta}]=\Pr[-\varepsilon/\sigma<\mathbf{x}'\boldsymbol{\beta}/\sigma]=\Phi(\mathbf{x}'\boldsymbol{\beta}/\sigma)$。于是,式(16.20)中的条件均值特别变成:

$$
\begin{aligned}
&\text{潜变量:} && \mathrm{E}[y^*|\mathbf{x}]=\mathbf{x}'\boldsymbol{\beta} \\
&\text{左截尾(在 0 点):} && \mathrm{E}[y|\mathbf{x},\ y>0]=\mathbf{x}'\boldsymbol{\beta}+\sigma\lambda(\mathbf{x}'\boldsymbol{\beta}/\sigma) \\
&\text{左删失(在 0 点):} && \mathrm{E}[y|\mathbf{x}]=\Phi(\mathbf{x}'\boldsymbol{\beta}/\sigma)\mathbf{x}'\boldsymbol{\beta}+\sigma\phi(\mathbf{x}'\boldsymbol{\beta}/\sigma)
\end{aligned}
\tag{16.23}
$$

类似地,可获得方差(参见 16.10.1 节)[1]。定义$w=\mathbf{x}'\boldsymbol{\beta}/\sigma$,我们得到:

$$
\begin{aligned}
&\text{潜变量:} && \mathrm{V}[y^*|\mathbf{x}]=\sigma^2 \\
&\text{左截尾(在 0 点):} && \mathrm{V}[y|\mathbf{x},\ y>0]=\sigma^2[1-w\lambda(w)-\lambda(w)^2] \\
&\text{左删失(在 0 点):} && \mathrm{V}[y|\mathbf{x}]=\sigma^2\Phi(w)\{w^2+w\lambda(w)+1-\Phi(w)[w+\lambda(w)]\}^2
\end{aligned}
\tag{16.24}
$$

很显然,截尾与删失引起异方差性,而对于截尾,$\mathrm{V}[y|\mathbf{x}]<\sigma^2$,因此如同人们所料,截尾缩减了可变性。

这些结果均假定正态误差。马达拉(Maddala, 1983,第 369 页)给出了关于对数正态分布、逻辑斯蒂分布、均匀分布、拉普拉斯分布、指数分布以及伽玛分布的类似于命题 16.1 的结论。

16.3.5 *Tobit 模型的边际效应*

边际效应是回归元上的变动对因变量条件均值的影响。这种效应随着关注内容是否在于潜变量均值$\mathbf{x}'\boldsymbol{\beta}$或由式(16.23)给出的截尾或删失均值而变化。一旦对每一个关于$\mathbf{x}$求微分,得到潜变量:

$$
\begin{aligned}
&\text{潜变量:} && \partial\mathrm{E}[y^*|\mathbf{x}]/\partial\mathbf{x}=\boldsymbol{\beta} \\
&\text{左截尾(在 0 点):} && \partial\mathrm{E}[y,\ y>0|\mathbf{x}]/\partial\mathbf{x}=\{1-w\lambda(w)-\lambda(w)^2\}\boldsymbol{\beta} \\
&\text{左删失(在 0 点):} && \partial\mathrm{E}[y|\mathbf{x}]/\partial\mathbf{x}=\Phi(w)\boldsymbol{\beta}
\end{aligned}
\tag{16.25}
$$

其中,$w=\mathbf{x}'\boldsymbol{\beta}/\sigma$,同时我们使用了$\partial\Phi(z)/\partial z=\phi(z)$以及$\partial\phi(z)/\partial z=-z\phi(z)$。删失均值的简单表达式,可通过某种处理来获得。它被分解成两种效应,其一是关于$y=0$的效应,其二是关于$y>0$的效应[参见麦克唐纳和莫菲特(McDonald and Moffitt, 1980)]。

在一些情况下,截尾或删失刚好是收集数据的人工制品,因此截尾与删失均值不是内在关注的内容,而我们对$\partial\mathrm{E}[y^*|\mathbf{x}]/\partial\mathbf{x}=\boldsymbol{\beta}$感兴趣。例如,就上端编码薪水数据而言,显然我们对测算受教育对平均薪水的影响而不是那些没有上端编码薪水的效应感兴趣。

在其他一些情况下,截尾或删失具有特定意义。例如,在工作小时模型中,式(16.25)中的三种边际效应分别对应于以下三种回归元变动效应:(1) 期望工作小

[1] 原著这里为习题 16.1,但应改为 16.10.1 节。——译者注

时;(2) 工人的实际工作小时;(3) 工人与非工人的实际工作小时。对于(1)来说,很明显,我们需要估计 $\boldsymbol{\beta}$,但对于(2)与(3)来说,显然关于 $\boldsymbol{\beta}$是非一致的,但斜率系数可能确实提供边际效应的一种合理粗略估计值,因为截尾与删失均值关于 $\mathbf{x}$ 仍然是线性的。

16.3.6 *Tobit 模型的可选择估计量*

除 MLE 之外,通过建立在关于截尾或删失均值正确表达式基础上的 NLS,可能获得一致估计。我们考察 NLS 估计量以及其他的最小二乘法估计量。

NLS 估计量

式(16.23)中的结果能用于通过 NLS 获得 Tobit 模型参数的一致估计值。例如,对于截尾数据,我们求

$$S_N(\boldsymbol{\beta},\sigma^2) = \sum_{i=1}^{N}(y_i - \mathbf{x}_i'\boldsymbol{\beta} - \sigma\lambda(\mathbf{x}_i'\boldsymbol{\beta}/\sigma))^2$$

既关于 $\boldsymbol{\beta}$ 的极小值,又关于 σ^2的极小值,然后实施对由式(16.24)给出的异方差性加以控制的推断。对于删失数据,可获得类似的估计值。

在实际应用中,并不使用这种估计量。一致性要求对截尾均值的正确设定,由式(16.21)知,这既需要误差的正态性,又需要同方差性。人们还可通过 ML 进行估计,因为这恰好依赖于强假设,同时是完全有效的。此外,在实际应用中,NLS 估计量可能是不精确的。由图 16.2 知,显然,$\lambda(\mathbf{x}'\boldsymbol{\beta}/\sigma)$关于 $\mathbf{x}'\boldsymbol{\beta}/\sigma$ 大致是线性的,由于 $\mathbf{x}$ 也是回归元,所以导致近似共线性。在 16.5 节,我们将考察,允许校正项类似于式(16.23)中的 $\sigma\lambda(\mathbf{x}'\boldsymbol{\beta}/\sigma)$模型,其优点是部分地依赖于回归元而不是 $\mathbf{x}$ 中的那些元素。

赫克曼两步估计量

由式(16.23)知,(在零点)截尾均值为:

$$\mathrm{E}[y|\mathbf{x}]=\mathbf{x}'\boldsymbol{\beta}+\sigma\lambda(\mathbf{x}'\boldsymbol{\beta}/\sigma) \tag{16.26}$$

若可以利用删失数据,则利用下述两步方法进行估计,而不使用 NLS。首先,对全部样本实施 d 对 $\mathbf{x}$ 的 probit 回归,当 $y>0$ 时是可观测的,二值变量 d 等于 0,从而得出一致估计值 $\hat{\boldsymbol{\alpha}}$,其中,$\boldsymbol{\alpha}=\boldsymbol{\beta}/\sigma$。其次,为了获得 $\boldsymbol{\beta}$ 与 σ 的一致估计值,对截尾样本实施 y 对 $\mathbf{x}$ 与 $\lambda(\mathbf{x}'\hat{\boldsymbol{\alpha}})$的 OLS 回归。

归功于赫克曼(Heckman, 1976, 1979)的这种估计方法,将在 16.5.4 节阐述,那里它将用于更一般的样本选择模型。16.10.2 节推导 $\hat{\boldsymbol{\beta}}$的标准误差,这可解释回归元 $\lambda(\mathbf{x}'\hat{\boldsymbol{\alpha}})$依赖于估计参数以及由截尾而引起的异方差性。

Tobit 模型的 OLS 估计

利用删失数据或截尾数据的 OLS 估计关于 $\boldsymbol{\beta}$ 是非一致的。这是因为由式(16.23)给出的删失与截尾均值并不等于 $\mathbf{x}'\boldsymbol{\beta}$,这违背了关于 OLS 一致性的根本条件。

对于删失数据,OLS 提供了对非线性删失回归曲线的一种线性近似。由图 16.1 与式(16.25)知,很明显,这条线与未删失数据的回归线相比更为平坦,其斜

率却等于真实斜率参数。戈德伯格(Goldberger, 1981)以解析形式已经证明,如果 y 与 $\mathbf{x}$ 是联合正态分布的,同时存在从下面 0 点进行删失,那么 OLS 斜率参数收敛到 p 倍的真实斜率参数,其中,p 表示具有正值样本部分。这些条件是约束性的,但鲁德(Ruud, 1986)却稍微放松了约束性。在实际应用中,若 Tobit 模型是适宜的,这种成比例结果提供了对 OLS 非一致性的良好经验近似。

类似地,具有截尾的回归线比未截尾回归线更为平坦。戈德伯格(Goldberger, 1981)得到了类似于删失情况的解析结果。如果 y 与 $\mathbf{x}$ 是联合正态分布的,且存在从下面 0 点进行的删失,那么 OLS 斜率参数收敛到真实参数的倍数。此倍数位于 0 与 1 之间,其表达式相当冗长。同时,其收缩程度与所有斜率系数一样。因此,截尾 OLS 低估了真实斜率参数的绝对数值。

16.3.7 *Tobit 模型的设定检验*

因 Tobit 模型具有脆弱性,一种好的实用做法是,检验分布是否被错误设定。存在四种广泛策略。

第一种方法是,在参数较为丰富的模型里面嵌套 Tobit 模型,并应用沃尔德、LR 或 LM 检验。由于零假设模型即 Tobit 模型是最容易进行估计的,自然是运用 LM 检验。尤其是,对删失回归模型中形式为 $\sigma_i^2=\exp(\mathbf{x}_i'\boldsymbol{\alpha})$ 的异方差性可直接进行检验。一旦利用 LM 检验的形式(参见 7.3.5 节),我们计算 N 次来自 1 对 $\tilde{\mathbf{s}}_{1i}$ 与 $\tilde{\mathbf{s}}_{2i}$ 的辅助回归的非中心 R^2,其中,$f_i=f(y_i|\mathbf{x}_i,\boldsymbol{\beta},\boldsymbol{\alpha})$ 表示由式(16.14)给出的密度,只是要用 $\exp(\mathbf{x}'\boldsymbol{\alpha})$ 代替 σ,关于 $\mathbf{s}_{1i}=\partial\ln f_i/\partial\boldsymbol{\beta}$ 与 $\mathbf{s}_{2i}=\partial\ln f_i/\partial\alpha$ 的表达式均可通过对式(16.16)中的表达式稍微修改来获得,而"~"表示在带有 $\boldsymbol{\alpha}$ 的所有分量的删失 Tobit MLE 处的计算值,除了截距等于 0 之外。对正态分布误差的假设进行检验的类似方法更加困难一些,因为不存在正态的标准一般化。

第二种方法是使用并不需要对备择假设模型设定的条件矩检验(参见 8.2 节)。特别地,关于删失 Tobit MLE 的一阶条件(16.16),建议基于广义残差:

$$e_i=d_i\frac{y_i-\mathbf{x}_i'\boldsymbol{\beta}}{\sigma^2}-(1-d_i)\frac{\phi_i}{\sigma(1-\Phi_i)}$$

的条件矩检验。若 Tobit 模型得以正确设定,则 $\mathrm{E}[e_i|\mathbf{x}_i]=0$,因为正规条件蕴含 $\mathrm{E}[\partial\ln f(y_i)/\partial\boldsymbol{\beta}]=0$。于是,我们可利用 $N^{-1}\sum_{i=1}^{N}\hat{e}_i\mathbf{z}_i$ 实施 H_0: $\mathrm{E}[e\mathbf{z}]=\mathbf{0}$ 与 H_a: $\mathrm{E}[e\mathbf{z}]\neq\mathbf{0}$ 检验,其中,$\hat{e}_i=e_i$ 表示在 Tobit MLE$(\hat{\boldsymbol{\beta}},\hat{\sigma}^2)$ 处计算。由 8.2.2 节,这种检验可通过计算 N 次来自 1 对 $\hat{e}_i\mathbf{z}_i$、$\hat{\mathbf{s}}_{1i}$ 以及 $\hat{s}_{2i}$ 的辅助回归的未中心 R^2,其中,$f_i=f(y_i|\mathbf{x}_i,\boldsymbol{\beta},\sigma^2)$ 表示由式(16.14)给出的密度,而由式(16.16)给出的 $\mathbf{s}_{1i}=\partial\ln f_i/\partial\boldsymbol{\beta}$ 与 $s_{2i}=\partial\ln f_i/\partial\sigma^2$,表示在$(\hat{\boldsymbol{\beta}},\hat{\sigma}^2)$处的计算值。$\mathbf{z}_i$ 变量可能是变量但不是 $\mathbf{x}_i$,在此情况下,该检验能被解释成对省略回归或 $\mathbf{x}_i$ 分量幂的检验。基于较高阶矩的条件矩检验同样可以得到发展。有关详细内容,参见切舍和艾里什(Chesher and Irish, 1987)以及帕甘和维拉(Pagan and Vella, 1989)。

第三种方法是,将为右删失持续期限数据而发展起来的某些诊断与检验方法(参见第 19 章),改写成左删失正态分布数据的情形。

最后一种方法是，把 β 的可供选择估计值与 Tobit MLE $\hat{\beta}$ 进行对比，这里可供选择的 β 估计值可以是 16.9 节阐述的著名半参数估计值，在较弱分布假设下，它是一致的。

对于进一步详细内容，参见帕甘和维拉(Pagan and Vella，1989)。他们阐述了带有某种应用的理论，而梅伦伯格和范·索斯特(Melenberg and Van Soest，1996)提供了更完整的应用。这两篇论文考察了除那些 Tobit 模型之外的较为丰富的样本选择模型的设定(参见 16.5 节)。

16.4 两部分模型

前面关于删失数据的一些模型，均将删失机制限制成与生成结果变量一样的模型。更一般地，删失机制与结果可利用独立的过程进行建模。例如，在解释个体每年度医疗费用支出时，第一种过程可决定住院治疗，而第二种过程可解释后来的医疗费用支出。要求两种独立机制的情况是强的，如果存在原因使人不得不认为，与较简单模型相比，某种实现值会以很大频率或者很小频率发生。例如，与同泊松分布相一致的情况相比，人们可能观测到更多的 0。允许 0 与非 0 由不同密度生成的两部分模型(two-part model)却增加了灵活性。实际上，这是混合模型的特定类型。

16.4.1 *两部分模型*

设具有完全可观测结果的个体为所研究活动的参与者(**participant**)。定义一个二值指示变量 d，对于参与者有 $d=1$，而对于非参与者有 $d=0$。假定对于参与者 $y>0$ 是可观测的，而对于非参与者 $d=0$ 是可观测的。对于非参与者，我们只能观测到 $\Pr[d=0]$。对于参与者，就某个选取的密度 $f(\cdot)$ 而言，给定 $y>0$ 时的条件密度被设定成 $f(y|d=1)$。于是，两部分模型由：

$$f(y|\mathbf{x})=\begin{cases}\Pr[d=0|\mathbf{x}], & \text{当 } y=0\\ \Pr[d=1|\mathbf{x}]f(y|d=1,\ \mathbf{x}), & \text{当 } y>0\end{cases} \tag{16.27}$$

给出。

此模型是由克拉格(Cragg，1971)详细阐述作为对 Tobit 模型的一般化，它可被表述成式(16.27)的一种特殊情况。参与决策 d 的明显模型是 probit 模型或 logit 模型。潜变量公式是，当 $I=\mathbf{x}'\boldsymbol{\beta}+\varepsilon$ 大于 0 时，$d=1$，而且该模型可被看成是一种围栏模型(**hurdle model**)，因为越过围绕栏或门限就变为参与者。为了确保参与者的正值，密度 $f(y|d=1,\ \mathbf{x})$ 应是正值随机变量，诸如对数正态或合适密度，例如，从下面 0 点正态截尾。

为了简单起见，通常两部分模型均会出现相同回归元，但这可以被放松，倘若存在明显的排除性约束，就应该得到放松。可直接实施极大似然估计，因为它分割成两种情形：利用所有观测值对离散选择模型进行估计，以及利用只满足 $y>0$ 的观测值对密度 $f(y|d=1,\ \mathbf{x})$ 参数进行估计。

16.4.2 *两部分模型例子*

端等人(Duan et al.，1983)利用源自兰德健康保险实验的数据，阐述了这种模型对预测医疗支出的重要应用。他们将一年期间是否有任何医疗支出设定成 probit 模型，因此 $\Pr[d=1|\mathbf{x}]=\Phi(\mathbf{x}_1'\boldsymbol{\beta}_1)$，而将给定某些支出的医疗设定成对数正态模型，因而 $\ln y|d=1,\mathbf{x}\sim\mathcal{N}[\mathbf{x}_2'\boldsymbol{\beta}_2,\sigma_2^2]$。于是，关于整个总体的期望医疗支出由

$$\mathrm{E}[y|\mathbf{x}]=\Phi(\mathbf{x}_1'\boldsymbol{\beta}_1)\exp[\sigma_2^2/2+\mathbf{x}_2'\boldsymbol{\beta}_2] \tag{16.28}$$

给出，其中第二项使用了下述结果：如果 $\ln y\sim\mathcal{N}[\mu,\sigma^2]$，那么 $\mathrm{E}[y]=\exp(\mu,\sigma^2/2)$。毛拉(Mullahy，1998)以更详细的方式考察了此类再变换。

就计数数据建模而言，两部分模型特别流行。例如，对医生出诊次数建模，存在一种模型决定病人是否看医生，而第二个模型决定那些至少已有一次看医生的病人后来看医生次数。然后，$\Pr[d=1]$ 被设定成泊松变量或负二项变量大于 0 的概率，而密度 $f(y|d=1)$ 被设定成从下面 0 点截尾的泊松密度或负二项密度。在计数文献中，归功于毛拉(Mullahy，1986)的这个模型称为围栏模型，将在 20.4.5 节中详述。

对于连续数据，两部分模型可用于含有过剩 0 的支出模型(克拉格的最初动机)。一种可供选择的样本选择模型在下一节阐述。

16.5 样本选择模型

在许多设置背景下都能产生样本选择，从而存在许多样本选择模型。在关注由赫克曼(Heckman，1979)所研究的二变量样本选择模型(**bivariate sample selection model**)重要例子之前，本节以对样本选择的一般讨论开始。另一个重要例子即罗伊模型(**Roy model**)将单独在 16.7 节加以研究。

16.5.1 *样本选择模型*

观测研究极少建立在纯随机样本上。更经常的方式是使用外生抽样(参见 3.2.4 节)，同时利用通常估计量。不过，若样本被有意或无意地部分建立在凭借因变量取值的基础上，则参数估计可能是非一致的，除非采用修正测量。这类样本被广泛定义成选择样本(**selection samples**)。·

由于存在许多方法生成选择样本，所以存在众多选择模型(**selection models**)。实际上，最容易被忽视的是，运用了选择样本。例如，考察当参与测验是自愿的时候，一段时期成绩测验例如 SAT 的平均分解释。最后时期可能归因于大学生知识的真实遗忘。不过，它或许刚好反映出选择效应：相对多的大学生参加一段时期测验，而新测验接收者则是相对很少的大学生。

选择可能归因于自选择(**self-selection**)，即关注结果部分地由个体是否选择参与到关注活动中而决定。它也可能起因于样本选择(**sample selection**)，即那些参与到关注活动中的个体者被故意过度抽样——极端情况是只抽取参与者。在上述两

种情况下，其中任何一种都会出现类似问题，选择模型通常被称为样本选择模型。

本章阐述文献中众多选择模型中的三种类型。一种最简单的模型是已在16.3节阐明的Tobit模型。一种普遍使用的典型模型，我们称之为二变量样本选择模型，将在本节余下部分加以阐述。通过引入不同于潜变量生成关注结果的删失潜变量，此种模型对Tobit模型加以推广。另一种流行的模型称为罗伊模型，将在16.7节阐述。该模型考察两个取值之一的结果，这样做要依赖于由删失随机变量所采用的值。这些模型分别对应于雨宫(Amemiya, 1985，第384页)的Tobit模型中第1、2、5类型。

在以不可观测因素为基础的样本选择情况下，一致估计依赖于相对强分布假设，甚至在半参数估计下也是如此。于是，实验数据研究提供了一种可供选择的引人注目的方法，因为选择问题可通过随机指派加以避免。不过，在经济应用中，出于成本及道德原因，很难实施实验。第25章将详述的处理效应方法试图将实验方法用于观测数据。

16.5.2 *二变量样本选择模型 (Tobit 模型第2类)*

设 y_2^* 表示关注的结果。在标准截尾Tobit模型中，当 $y_2^*>0$ 时，该结果是可观测的。更一般的模型更要引入不同的潜变量 y_1^*，并且当 $y_1^*>0$ 时，此结果 y_2^* 是可观测的。例如，y_1^* 决定是否去工作，而 y_2^* 决定工作多少个小时，同时 $y_1^*\neq y_2^*$，因为去工作存在固定成本，诸如交通成本，而一旦去工作，交通成本在确定参加工作与否方面比工作多少个小时更为重要。

二变量样本选择模型(**bivariate sample selection model**)包括参与方程(**participation equation**)，即：

$$y_1=\begin{cases}1, & \text{当 } y_1^*>0\\ 0, & \text{当 } y_1^*\leqslant 0\end{cases} \tag{16.29}$$

以及相应的结果方程(**participation equation**)：

$$y_2=\begin{cases}y_2^*, & \text{当 } y_1^*>0\\ -, & \text{当 } y_1^*\leqslant 0\end{cases} \tag{16.30}$$

这个模型设定，当 $y_1^*>0$ 时，y_2 是可观测的，而当 $y_1^*\leqslant 0$ 时，y_2 不需要取任何有意义的值。其标准模型设定成线性模型带有潜变量的可加误差形式，因而：

$$\begin{aligned}y_1^*&=\mathbf{x}_1'\boldsymbol{\beta}_1+\varepsilon_1\\ y_2^*&=\mathbf{x}_2'\boldsymbol{\beta}_2+\varepsilon_2\end{aligned} \tag{16.31}$$

如果 ε_1 与 ε_2 是相关的，那么估计 $\boldsymbol{\beta}_2$ 时便会出现问题。显然，Tobit模型是 $y_1^*=y_2^*$ 的一种特殊情况。

此模型并不存在广泛接受的称谓。赫克曼(Heckman, 1979)使用它阐明给定样本选择时的估计问题。这一模型等价于带有随机门限的Tobit模型[纳尔逊(Nelson, 1977)]。假定当 $y_2^*>L^*$ 时，我们可观测到 y_2^*，其中，y_2^* 如同式(16.31)所定义的，同时门限是 $L^*=\mathbf{z}'\boldsymbol{\gamma}+v$，而不是16.3节中的 $L^*=0$。于是，等价地当

$y_1^*>0$ 时，我们可观测到 y_2^*，其中，$y_1^*=y_2^*-L^*=(\mathbf{x}_2'\boldsymbol{\beta}_2-\mathbf{z}'\boldsymbol{\gamma})+(\varepsilon_2-v)=\mathbf{x}_1'\boldsymbol{\beta}_1+\varepsilon_1$，而 $\mathbf{x}_1$ 表示 $\mathbf{x}_2$ 与 $\mathbf{z}$ 的并，同时 $\boldsymbol{\beta}_1$ 与 ε_1 均以明显方式加以定义。雨宫(Amemiya, 1985，第 384 页)称此模型为 Tobit 模型第二类。伍德里奇(Wooldridge, 2002，第 506 页)将此模型称为含有 probit 选择方程的 Tobit 模型。尽管存在如此之多的这类模型，但其他一些学者则称此模型为广义模型或样本选择模型。

给定另一个假设：相关误差服从联合正态分布且同方差，并满足：

$$\begin{bmatrix}\varepsilon_1\\ \varepsilon_2\end{bmatrix}\sim\mathcal{N}\left[\begin{bmatrix}0\\ 0\end{bmatrix},\begin{bmatrix}1 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{bmatrix}\right] \tag{16.32}$$

则可通过 ML 直接进行估计。至于 14.4.1 节的 probit 模型，由于 y_1^* 的唯一符号是可观测的，所以可使用正规化 $\sigma_1^2=1$。

已知式(16.29)与式(16.30)，对于 $y_1^*>0$，我们以下述概率可以观测到 y_2^*，此概率等于 $y_1^*>0$ 的概率乘以给定 $y_1^*>0$ 时 y_2^* 的条件概率。因而，对于正的 y_2，可观测的密度为 $f^*(y_2^*|y_1^*>0)\times\Pr[y_1^*>0]$。对于 $y_1^*\leqslant 0$，可观测的所有内容是这个事件发生，而且密度是这一事件发生的概率。因此，二变量样本选择模型具有似然函数：

$$\mathrm{L}=\prod_{i=1}^{n}\{\Pr[y_{1i}^*\leqslant 0]\}^{1-y_{1i}}\{f(y_{2i}|y_{1i}^*>0)\times\Pr[y_{1i}^*>0]\}^{y_{1i}} \tag{16.33}$$

其中，当 $y_{1i}^*\leqslant 0$ 时第一项表示离散分布，从而 $y_{1i}=0$，而当 $y_{1i}^*>0$ 时第二项表示连续分布。这一似然函数可应用于相当一般模型，而不只是含有联合正态误差的线性模型。

对含有联合正态误差的线性模型进行专门化研究，会得到作为正态的二变量密度 $f^*(y_1^*, y_2^*)$，导致第二项中的条件密度成为单变量正态的，而且很容易加以处理。雨宫(Amemiya, 1985，第 385～387 页)曾经提供详细内容，包括似然函数的准确形式。

这种模型的早期经典应用是劳动力供给，其中，y_1^* 表示不可观测的意愿或者工作倾向，而 y_2 表示实际工作小时。与 14.2.1 节中需要工作"意愿"小时技巧的 Tobit 模型相比，此模型在概念上对劳动力供给更具有吸引力。这种典型应用会具有以下复杂情况，即对于那些不参加工作的个体来说，重要回归元即工资报价的数据出现缺失。于是，严格地说，尽管该模型不只是二变量样本选择模型，但其复杂情况可通过添加工资报价方程，并代入其中来加以应对。关于对劳动力供给的出色应用的内容，参见姆罗茨(Mroz, 1987)。

16.5.3 *二变量样本选择模型的条件均值*

在本节，我们要获得二变量样本选择模型的条件截尾均值。它跟 $\mathbf{x}_2'\boldsymbol{\beta}_2$ 不一样，因而 y_2 对 $\mathbf{x}_2$ 的 OLS 回归会导致非一致参数估计。不过，条件均值表达式却能用于激发一种可供选择的估计方法，与 MLE 所需要的分布假设相比，该方法依赖于较弱分布假设。这将在下一节给出。

我们考察样本选择模型中的截尾均值，这里只有 y_2 的正值可以利用。通常，这是：

$$E[y_2 \mid \mathbf{x}, y_1^* > 0] = E[\mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon_2 \mid \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1 > 0] \qquad (16.34)$$
$$= \mathbf{x}_2'\boldsymbol{\beta}_2 + E[\varepsilon_2 \mid \varepsilon_1 > -\mathbf{x}_1'\boldsymbol{\beta}_1]$$

其中,$\mathbf{x}$ 表示 $\mathbf{x}_1$ 与 $\mathbf{x}_2$ 的并。如果误差 ε_1 与 ε_2 是独立的,那么最后一项简化成 $E[\varepsilon_2]=0$,同时 y_2 对 $\mathbf{x}_2$ 的 OLS 回归将会得出 $\boldsymbol{\beta}_2$ 的一致估计。然而,两个误差项之间的任何相关性意味着截尾均值不再是 $\mathbf{x}_2'\boldsymbol{\beta}_2$,可是我们需要对选择加以解释。

当 ε_1 与 ε_2 是相关的时,为了获得 $E[\varepsilon_2 \mid \varepsilon_1 > -\mathbf{x}_1'\boldsymbol{\beta}_1]$,赫克曼(Heckman,1979)注意到,若式(16.31)中的误差(ε_1, ε_2)如同式(16.32)一样是联合正态的,则下述式(16.36)蕴含:

$$\varepsilon_2 = \sigma_{12}\varepsilon_1 + \xi \qquad (16.35)$$

其中,随机变量 ξ 与 ε_1 是独立的。为了得出这个结果,注意,通常联合正态分布:

$$\begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{bmatrix} \sim \mathcal{N}\left[\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right]$$

蕴含着条件正态分布:

$$\mathbf{z}_2 \mid \mathbf{z}_1 \sim \mathcal{N}[\boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{z}_1 - \boldsymbol{\mu}_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}]$$

该结果意味着:

$$\mathbf{z}_2 = \boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{z}_1 - \boldsymbol{\mu}_1) + \boldsymbol{\xi} \qquad (16.36)$$

其中,$\boldsymbol{\xi} \sim \mathcal{N}[\mathbf{0}, \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}]$ 与 $\mathbf{z}_1$ 是独立的。对于由式(16.32)给出的联合密度,我们具有纯量形式 $\mu_1 = \mu_2 = 0$ 且 $\sigma_1^2 = 1$,所以式(16.36)专门化为式(16.35)。

通过利用式(16.35),截尾(16.34)变成:

$$E[y_2 \mid \mathbf{x}, y_1^* > 0] = \mathbf{x}_2'\boldsymbol{\beta}_2 + E[(\sigma_{12}\varepsilon_1 + \xi) \mid \varepsilon_1 > -\mathbf{x}_1'\boldsymbol{\beta}_1]$$
$$= \mathbf{x}_2'\boldsymbol{\beta}_2 + \sigma_{12}E[\varepsilon_1 \mid \varepsilon_1 > -\mathbf{x}_1'\boldsymbol{\beta}_1]$$

其中,我们使用了 ξ 与 ε_1 的独立性。选择项类似于较简单的 Tobit 模型中的情况,而且再一次使用命题 16.1 中的 $E[z \mid z > -c]$ 的表达式,我们得到:

$$E[y_2 \mid \mathbf{x}, y_1^* > 0] = \mathbf{x}_2'\boldsymbol{\beta}_2 + \sigma_{12}\lambda(\mathbf{x}_1'\boldsymbol{\beta}_1) \qquad (16.37)$$

其中,$\lambda(z) = \phi(z)/\Phi(z)$,并且我们使用了 $\sigma_1^2 = 1$。类似地,命题 16.1(iii)会得出截尾方法:

$$V[y_2 \mid \mathbf{x}, y_1^* > 0] = \sigma_2^2 - \sigma_{12}^2\lambda(\mathbf{x}_1'\boldsymbol{\beta}_1)(\mathbf{x}_1'\boldsymbol{\beta}_1 + \lambda(\mathbf{x}_1'\boldsymbol{\beta}_1)) \qquad (16.38)$$

当 $y_1^* \leqslant 0$ 时,前面分析没有设定值。在一些应用中,当 $y_1^* < 0$ 时,y_2 可能等于 0。于是,考察删失均值有意义。把可观测的 y_2 和不可观测的 y_1^* 及 y_2^* 作为条件,然后进行无条件化处理,得到:

$$\begin{aligned} E[y_2 \mid \mathbf{x}] &= E_{y_1^*}E[y_2 \mid \mathbf{x}, y_1^*] \qquad (16.39) \\ &= \Pr[y_1^* \leqslant 0 \mid \mathbf{x}] \times 0 + \Pr[y_1^* > 0 \mid \mathbf{x} \times E[y_2^* \mid \mathbf{x}, y_1^* > 0] \\ &= 0 + \Phi(\mathbf{x}_1'\boldsymbol{\beta}_1)\{\mathbf{x}_2'\boldsymbol{\beta}_2 + \sigma_{12}\lambda(\mathbf{x}_1'\boldsymbol{\beta}_1)\} \\ &= \Phi(\mathbf{x}_1'\boldsymbol{\beta}_1)\mathbf{x}_2'\boldsymbol{\beta}_2 + \sigma_{12}\phi(\mathbf{x}_1'\boldsymbol{\beta}_1) \end{aligned}$$

其中,第三行使用式(16.37),而最后一行使用 $\lambda(z) = \phi(z)/\Phi(z)$。可以证明,删失方差是异方差的。

16.5.4 赫克曼两步估计量

y_2 对 $\mathbf{x}_2$ 的 OLS 回归仅仅利用了可观测的 y_2 正值，这个重要的结果会导致 $\boldsymbol{\beta}$ 的非一致估计，除非误差是无关的，因而 $\sigma_{12}=0$。很明显，由截尾均值公式(16.37)知，这里还包括了“回归元”$\lambda(\mathbf{x}_1'\boldsymbol{\beta}_1)$。

赫克曼两步法(**Heckman's two-step procedure**)有时称为 Heckit 估计量，它借助于省略回归元 $\lambda(\mathbf{x}_1'\boldsymbol{\beta}_1)$ 的估计值增大 OLS 回归。因而，利用 y_2 的正值，通过 OLS 对模型

$$y_{2i}=\mathbf{x}_{2i}'\boldsymbol{\beta}_2+\sigma_{12}\lambda(\mathbf{x}_{1i}'\hat{\boldsymbol{\beta}}_1)+v_i \tag{16.40}$$

进行估计，其中，v 表示误差项，$\hat{\boldsymbol{\beta}}_1$ 可通过 y_1 对 $\mathbf{x}_1$ 的第一步回归来获得，因为 $\Pr[y_1^*>0]=\Phi(\mathbf{x}_1'\boldsymbol{\beta}_1)$，而 $\lambda(\mathbf{x}_1'\hat{\boldsymbol{\beta}}_1)=\phi(\mathbf{x}_1'\hat{\boldsymbol{\beta}}_1)/\Phi(\mathbf{x}_1'\hat{\boldsymbol{\beta}}_1)$ 表示估计的逆米尔斯比值。该回归没有直接提供 σ_2^2 的估计值，但截尾方差公式(16.38)产生了 $\hat{\sigma}_2^2=N^{-1}\sum_i\times[\hat{v}_i^2+\hat{\sigma}_{12}^2\hat{\lambda}_i(\mathbf{x}_1'\hat{\boldsymbol{\beta}}_1+\hat{\lambda}_i)]$ 估计值，其中，$\hat{v}_i$ 源自式(16.40)的 OLS 残差，而 $\hat{\lambda}_i=\lambda(\mathbf{x}_{1i}'\hat{\boldsymbol{\beta}}_1)$。然后，式(16.32)中的两个误差之间的相关性可通过 $\hat{\rho}=\hat{\sigma}_{12}/\hat{\sigma}_2$ 得到估计。

对 $\sigma_{12}=0$ 或 $\rho=0$ 是否成立进行检验，就是对误差是否相关进行检验，而且需要样本选择相关性。这类检验是建立在 $\hat{\sigma}_{12}$ 基础上的沃尔德检验，$\hat{\sigma}_{12}$ 表示逆米尔斯比的估计系数。

重要的是注意到，由回归(16.40)报告的通常 OLS 标准误差是不正确的，异方差稳健标准误差也是不正确的。标准误差正确公式考虑到了两阶段回归中的两个复杂情况。第一，即使 $\boldsymbol{\beta}_1$ 是已知的，但式(16.40)中的误差是源自式(16.38)的异方差。第二，实际上，$\boldsymbol{\beta}_1$ 可用其估计值来代替，对于较简单的 Tobit 模型，6.6 节也研究了复杂情况，而 16.10.2 节则进行了分析。正确标准误差公式是由赫克曼(Heckman, 1979)给出的；还可参见格林(Greene, 1981)。16.10.2 节推导出较简单 Tobit 模型的这些公式。实施起来并不简单，因而最好是使用可以自助处理这种复杂情况的软件包或运用自助法。

所得到的 $\boldsymbol{\beta}_2$ 估计量是一致的。在误差联合正态性下，与 MLE 相比，虽然有效性损失是相当大的，可是因为下述原因，该估计量颇为流行：(1) 它实施起来简单；(2) 此方法适用于一系列的选择模型，包括由 16.7 节给出的那些模型；(3) 该估计量需要比 ε_1 与 ε_2 联合正态性更弱的分布假设；(4)为了进一步允许如同 16.9 节一样的半参数估计，甚至可对这些分布假设加以削弱。

所需要的关键假设(**key assumption**)是式(16.35)，本质上有：

$$\varepsilon_2=\delta\varepsilon_1+\xi \tag{16.41}$$

其中，ξ 与 ε_1 是独立的。这看起来是相当合理的。在耐用品支出情况下，比如说，该式表明支出方程式中的误差是购买决策方程中误差的多倍，加上与购买决策独立的某个噪声；实质上为关于误差的线性回归模型。已知假设(16.41)，条件均值(16.34)变成：

$$\mathrm{E}[y_2|y_1^*>0]=\mathbf{x}_2'\boldsymbol{\beta}_2+\delta\mathrm{E}[\varepsilon_1|\varepsilon_1>-\mathbf{x}_1'\boldsymbol{\beta}_1] \tag{16.42}$$

若 ε_1 服从标准正态分布，则这会产生式(16.37)，即 OLS 回归(16.40)的基础。

更一般地，赫克曼两步方法能用于含有 ε_1 分布而不是正态情形的式(16.42)；例如，参见奥尔森(Olsen, 1980)。人们还能运用不施加关于 $\mathrm{E}[\varepsilon_1 | \varepsilon_1 > -\mathbf{x}_1'\boldsymbol{\beta}_1]$ 的函数形式的半参数方法(参见 16.9 节)。

16.5.5 *识别考虑*

若对回归元没有任何约束，含有正态误差的二变量样本选择模型在理论上是可识别的。特别地，完全一样的回归元能出现在关于 y_1^* 与 y_2^* 的方程中。

然而，如果使用完全相同的回归元，那么具有正态分布误差的模型接近于不可识别的。若 $\mathbf{x}_1 = \mathbf{x}_2$，则 $\mathrm{E}[y_2 | y_1^* > 0] \simeq \mathbf{x}_2'\boldsymbol{\beta}_2 + a + b\mathbf{x}_2'\boldsymbol{\beta}_1$，一旦利用式(16.37)与源自 16.3.2 节的观测值，逆米尔斯比项 $\lambda(\cdot)$ 在它的自变量广泛区域大致为线性的。这就产生显而易见的多重共线性问题，许多文章对此进行了讨论，包括绳田(Nawata, 1993)、绳田和长濑(Nawata and Nagase, 1996)，以及梁和余(Leung and Yu, 1996)。利用 10.4.2 节给出的条件数，可以发现多重共线性，由式(16.40)知，回归元是 $\mathbf{x}_2$ 以及 $\lambda(\mathbf{x}_1'\hat{\boldsymbol{\beta}}_1)$。对于不同观测值而言，$\mathbf{x}_1'\hat{\boldsymbol{\beta}}_1$ 上的变异较大时，则问题就不算严重，也就是说，较好的 probit 模型能在参与及非参与之间进行辨别。

赫克曼两步法的一些半参数变形(参见 16.9.3 节)确实需要排除性约束。因此，具有正态误差的二变量样本选择模型的识别，可通过函数形式假设来达到。

因而，应用中，对二变量样本选择模型加以估计可能需要参与方程(y_1^*)中的至少一个回归元被排除在结果方程(y_2^*)之外。例如，与工作小时数无关的工作固定成本将会影响到是否工作的决策，但不会影响工作小时数。如同许多应用一样，比如 16.6 节，这是一个很大的局限性，做出保护性的排除约束是相当难的。

16.5.6 *边际效应*

二变量样本选择模型的边际效应会依照我们是否考虑潜变量均值或由式(16.37)给出的截尾均值或者删失均值(如果它是合适的)而变化。

将 $\mathbf{x}$ 定义成由 $\mathbf{x}_1$ 与 $\mathbf{x}_2$ 的并形成的一个向量，同时将 $\mathbf{x}_1'\boldsymbol{\beta}_1$ 重新写成 $\mathbf{x}'\boldsymbol{\gamma}_1$，而将 $\mathbf{x}_2'\boldsymbol{\beta}_2$ 重新写成 $\mathbf{x}'\boldsymbol{\gamma}_2$，这样做会很方便。例如，截尾均值变成 $\mathrm{E}[y_2 | \mathbf{x}] = \mathbf{x}'\boldsymbol{\gamma}_2 + \sigma_{12}\lambda(\mathbf{x}'\boldsymbol{\gamma}_1)$。注意到，如果 $\mathbf{x}_1 \neq \mathbf{x}_2$，那么 $\boldsymbol{\gamma}_1$ 与/或 $\boldsymbol{\gamma}_2$ 将拥有某些零元素。对 $\mathbf{x}$ 进行微分，得到删失：

$$
\begin{aligned}
&\text{未删失：} && \partial \mathrm{E}[y_2^* | \mathbf{x}]/\partial \mathbf{x} = \boldsymbol{\gamma}_2 \qquad (16.43)\\
&\text{截尾(在 0 点)：} && \partial \mathrm{E}[y_2 | \mathbf{x}, y_1 = 1]/\partial \mathbf{x} = \boldsymbol{\gamma}_2 - \sigma_{12}\lambda(\mathbf{x}'\boldsymbol{\gamma}_1)(\mathbf{x}'\boldsymbol{\gamma}_1 + \lambda(\mathbf{x}'\boldsymbol{\gamma}_1))\\
&\text{删失(在 0 点)：} && \partial \mathrm{E}[y_2 | \mathbf{x}]/\partial \mathbf{x} = \boldsymbol{\gamma}_1 \phi(\mathbf{x}'\boldsymbol{\gamma}_1)\mathbf{x}'\boldsymbol{\gamma}_2 + \Phi(\mathbf{x}'\boldsymbol{\gamma}_1)\boldsymbol{\gamma}_2 - \sigma_{12}\mathbf{x}'\boldsymbol{\gamma}_1 \phi(\mathbf{x}'\boldsymbol{\gamma}_1)\boldsymbol{\gamma}_1
\end{aligned}
$$

其中，$\lambda(z) = \phi(z)/\Phi(z)$，并且我们使用 $\partial\phi(z)/\partial z = -z\phi(z)$，以及 $\partial\lambda(z)/\partial z = -z\phi(z)/\Phi(z) - \phi(z)^2/\Phi(z)^2 = -\lambda(z)(z + \lambda(z))$。对这三个导数的解释，可类似于 16.3.5 节以某种详细方式所做出的讨论。如同已经提及的，当 $y_1 = 0$，只有 y_2 取 0 值时，才适合于进行删失均值的分析。在一些应用中，诸如稍后将要讨论的健康支出的自然对数，没有删失均值。

16.5.7 *依可观测因素与不可观测因素的选择*

存在如下一些建模情形：将建模考虑成两部分决策问题：首先是参加活动，然后决定活动的水平。这些决策缠绕在一起，同时可认为是依赖于共同因素。此类数据的一个自然模型是，二变量选择模型(16.29)～(16.31)。

在一些情况下，包括回归元以后，两个过程中任何剩下的误差(ε_1 与 ε_2)是不相关的。例如，对于住院治疗模型，可能会是在一旦控制了观测个体特征，诸如健康状况后，在决定入院医治方程的误差与决定住院多久方程的误差之间不存在相关性。在那种情况下，可直接进行分析，因为选择仅仅是建立在可观测因素的基础上，例如当 $\sigma_{12}=0$ 时式(16.37)可以简化。这两部分能独自建模，同时使用 16.4 节中较简单的两部分模型。

在另外一些情况下，甚至包括回归元以后，误差可能是相当的。例如，在劳动力供给中，促使某人可能去工作的不可观测因素，也可能导致他们与通过可观测回归元所预测的工作小时数相比，可能工作小时数更长。人们能检验误差之间是否存在此类相关。若存在相关，则选择以不可观测因素进行，从而本章的方法开始发挥作用。甚至对于赫克曼两步法，需要相对强的分布假设。

由端等人(Duan et al., 1983)做出的研究已在 16.4.2 节概述，因为运用了比样本选择模型更为引人注目的两部分模型而受到批评。这导致了激烈的争论，大多数有关文章均列在梁和余(Leung and Yu, 1993)的参考文献里，他们突出了逆米尔斯比值与剩余回归元之间潜在关系的重要作用。

更一般地，一些选择模型诸如两变量选择模型，既允许依据可观测因素选择又允许依据不可观测因素选择，因为它既依据可观测回归元选择，又依据不可观测误差选择。就依据隐性可观测的选择而言，更简单地称为依据不可测因素选择(**selection on unobservables**)的模型。本章强调依据不可观测因素的选择。

相反，如果我们只依据可观测因素选择(**selection on observables**)，那么分析变得更为简单。本章的两部分模型就是一个例子。关于处理评估的第 25 章，强调依据可观测因素的选择(参见 25.3.3 节的讨论)，同时详述了诸如倾向得分匹配方法。

16.6 选择例子：健康支出

为阐述方便，我们使用源自兰德健康保险实验(RHIE)的数据。节选数据来自德布和特里维蒂(Deb and Trivedi, 2002)，他们对去看内科大夫的门诊病人数以及所有提供者进行建模，所用模型为计数模型。20.3 节归纳了这些数据，而 20.7 节阐述了一些标准计数模型的估计。

然而，我们这里对每年健康支出进行建模。回归元与表 20.4 详细定义的回归元一样。将它们分成健康保险变量(LC, IDP, LPI 和 FMDE)、社会经济特征(LINC, LFAM, AGF, FEMALE, CHILD, FEMCHILD, BLACK 和 EDUCDEC)，以及健康状况变量(PHYSLIM, NDISEASE, HLTHF 和 HLTHP)。第 20 章的分析使用 4 年的数据，而我们这里仅使用 2 年的数据，得到 5 574 个观测值，并概括

统计量，这些概括统计量与表 20.4 给出的那些并不完全一样。

因变量 y 表示年度个体健康支出。经济计量模型要考虑两种复杂情况：(1)健康支出为 0 的占样本 23.2%；(2)正的健康支出具有非常向右的偏斜度，其均值为 221 美元，该值远远大于中位数 53 美元。对数变换可剔除这种偏斜度，所得均值 4.07 接近于中位数 3.96，从而偏斜度位于 24.0 到 0.3。其峰度为 3.29，接近于正态值 3。

我们关注于那些正医疗支出的 $\ln y$ 建模。一些可行模型包括两部分模型，对 16.4.2 节的医疗支出对数以及二变量样本选择模型(参见 16.5.2 节)加以阐明，其中，式(16.29)中的 y_1 表示正支出的指示变量，而式(16.30)中的 y_2 表示 $\ln y$。注意到，考察当 $y_1=0$ 时，y_2 的值是没有意义的，因为 $\ln 0$ 没有定义。两部分模型是满足式(16.32)中 $\sigma_{12}=0$ 的二变量样本选择模型的特殊情况。

表 16.1 给出健康保险变量与健康状况回归元的一些结果。为了简单起见，回归中同样包含的社会经济变量从该表中省略了。

表 16.1 健康支出数据：来自两部分模型与选择模型[a]

模型	两部分		选择两部分		选择 MLE	
方程	DMED	LNMED	DMED	LNMED	DMED	LNMED
LC	−0.119	−0.016	−0.119	−0.028	−0.107	−0.076
	(−4.41)	(−0.52)	(−4.41)	(−0.70)	(−4.03)	(2.25)
IDP	−0.128	−0.079	−0.128	−0.028	−0.109	−0.150
	(−2.45)	(−1.28)	(−2.45)	(−0.70)	(−2.13)	(−2.26)
LPI	−0.028	0.003	0.028	0.005	0.029	0.015
	(3.19)	(0.28)	(3.19)	(0.47)	(3.42)	(1.42)
FMDE	0.008	−0.031	0.008	−0.030	0.001	−0.024
	(0.47)	(−1.69)	(0.47)	(−1.62)	(0.05)	(1.21)
PHYSLIM	0.273	0.262	0.273	0.281	0.285	0.355
	(3.67)	(3.81)	(3.67)	(3.50)	(3.94)	(4.70)
NDISEASE	0.022	0.022	0.022	0.022	0.021	0.029
	(6.25)	(5.78)	(6.25)	(4.29)	(6.03)	(7.54)
HLTHG	0.039	0.144	0.039	0.147	0.058	0.156
	(0.88)	(2.94)	(0.88)	(3.01)	(1.35)	(2.99)
HLTHF	0.192	0.364	0.192	0.382	0.224	0.445
	(2.29)	(4.13)	(2.29)	(3.98)	(2.75)	(4.66)
HLTHP	0.640	0.787	0.640	0.833	0.798	0.999
	(3.01)	(4.63)	(3.01)	(4.22)	(3.90)	(5.32)
ρ		0.000		0.168		0.736
σ_2				1.401		1.570
$\sigma_{12}=\sigma_2\rho$		0.000		0.236		1.155
				(0.47)		(16.43)
$-\ln L$	10 184.1				10 170.1	

[a] t 统计量位于括号中。回归元还包括了 8 个社会经济特征。DMED 表示医疗支出是否为正的指示变量，而 LNMED 表示支出的自然对数，假如支出为正的话。两步选择模型的第二步 t 统计量是建立在以下误差基础上，该误差对用于获得拟合逆米尔斯比值项的第一步加以校正。

首先，我们将两部分模型估计值与二变量样本选择模型的两步估计值加以比较。DMED 方程估计值，与通过 DMED 对同样回归元做出的 probit 回归所获得的那些值一样。LNMED 方程的估计值就不同，因为关于 LNMED 的两步样本选择的回归，第二步还包括了逆米尔斯比率项回归元。这个额外项是统计不显著的（$t=0.47$），且数值很小，得出 $\hat{\rho}=0.168$，接近于 0。因此，两个模型得出的 LNMED 方程系数估计值相似。

正如 16.4.4 节所提及的，当逆米尔斯比率项与其他回归元高度相关时，两步估计量执行效果就不好。这里，没有出现这种情况，因为 probit 模型预测概率存在范围从 0.15 到 0.99，且在第二阶段中第二阶段回归的条件数目（参见 10.4.4 节）虽然有点大，但通过包括逆米尔斯比率仅仅增加一倍，即从 37 到 82。尽管人们拥有某些排除性约束仍然是更可取的，但在此应用中，DMED 中的哪些回归元建立在 LNMED 方程的先验基础上被合理地排除掉并不清楚。

不论是 DMED 方程，还是 LNMED 方程，二变量样本选择模型的 ML 估计值，截然不同于前面的估计值。DMED 与 LNMED 的潜变量模型中的误差是高度相关的，估计值 $\hat{\rho}=0.736$，这是非常统计显著的（$t=16.43$）。σ_{12}（或者 ρ）的两步估计值与 ML 估计值之间的巨大差异，最好被认为是，显示二变量样本选择模型存在问题的信号。对零假设——估计值具有相同的概率极限——拒绝，可利用 8.4 节给出的豪斯曼检验，能被解释成对两变量选择模型从两步估计到 ML 估计所需的另外联合正态性假设的拒绝。不过，或许存在更为基本的问题，满足较弱假设（16.41）与 ε_1 iid 正态的二变量样本选择模型同样是不合理的。二变量样本选择模型的这种脆弱性并不异乎寻常，尤其是如果该模型的两部分均使用相同回归元，那么通过模型设定假设可实现识别。此处，它是通过利用健康支出数据合成的，这些支出数据具有相当大的离群值[1]（outliers），因此误差可能不是正态的，即使 LNEED 具有接近于 0 的偏斜度，且峰度接近于 3，正如已经提及的，对异方差性、偏斜度以及峰度的标准检验完全拒绝（p 值为 0.000 0）零假设：LNMED 是正态分布的。

最受关注的回归元是 LC，即共保险率的自然对数，共保险率是由病人支付投保的健康成本百分比形成的。最为统计显著的效应是决定支出是否为正的，而不是正支出的大小。若所有观测值均是正的，则关于 LNMED 回归中的 LC 系数等于对健康保健需求的价格弹性。实际上，在预测支出对数的条件截尾均值的价格变动影响时，我们要求控制那些支出为 0 的效应，如同式（16.43）的第二行一样。

在一些应用中，关注内容在于预测，而不是对边际效应进行估计。在该例子中，对于想要预测支出水平而不是支出对数，这是复杂情况。一旦假定对数正态性，两部分模型的表示是由式（16.28）给出的。端等人（Duan et al.，1983）阐述了，在没有对数正态假设下进行预测的方法，可看成是自助法的一种变形。也可参见毛拉（Mullahy，1998）。

---

〔1〕 又称为异常值。——译者注

16.7 罗伊模型

在二变量样本选择模型中，关于个体的因变量可能是不可观测的。因此，当 $y_1=1$ 时，就个体而言，我们观测到 y_2，但当 $y_1=0$ 时，则根本观测不到 y_2。在本节，我们考察对所有个体而言可观测到 y_2 的那种模型，但只为两种可能状态之一。此类重要模型强调反事实框架(**counterfactuals**)，并与第 25 章阐述的项目评估文献有联系。

16.7.1 *罗伊模型*

被罗伊(Ray，1951)经常引用的文章，对当个体技能存在异质性以及个体寻找职业自我选择时工资的职业分布结果(既有均值又有方差)加以考察。尽管假定个体工人的职业产出在没有选择的条件下是对数正态的，同时根本不考察正式模型的估计，但其研究相对来说，则是一般性的且并不精准。在 20 世纪 70 年代，许多学者独立提出利用横截面数据加以估计的类似情况，并考察既依据可观测因素选择又依据不可观测因素选择。这类模型即为著名的罗伊模型。

我们将原形罗伊模型(**Roy model**)定义如下。潜变量 y_1^* 决定观测到的结果是否是 y_2^* 或 y_3^*。具体地讲，我们观测到 y_1^* 为正或为负：

$$y_1=\begin{cases}1, & \text{当 } y_1^*>0\\ 0, & \text{当 } y_1^*\leqslant 0\end{cases}\tag{16.44}$$

并且依据：

$$y=\begin{cases}y_2^*, & \text{当 } y_1^*>0\\ y_3^*, & \text{当 } y_1^*\leqslant 0\end{cases}\tag{16.45}$$

准确地观测到 y_2^* 与 y_3^* 之一。

一种习惯做法是，设定关于潜变量的线性模型，且具有可加误差，满足：

$$\begin{aligned}y_1^*&=\mathbf{x}_1'\boldsymbol{\beta}_1+\varepsilon_1\\ y_2^*&=\mathbf{x}_2'\boldsymbol{\beta}_2+\varepsilon_2\\ y_3^*&=\mathbf{x}_3'\boldsymbol{\beta}_3+\varepsilon_3\end{aligned}\tag{16.46}$$

具有可加效应的模型是设定 $\mathbf{x}_3'\boldsymbol{\beta}_3=\mathbf{x}_2'\boldsymbol{\beta}_2+\alpha$ 形式。关于相关误差的最简单参数模型是联合正态的，满足：

$$\begin{bmatrix}\varepsilon_1\\ \varepsilon_2\\ \varepsilon_3\end{bmatrix}\sim\mathcal{N}\left[\begin{bmatrix}0\\ 0\\ 0\end{bmatrix},\begin{bmatrix}1 & \sigma_{12} & \sigma_{13}\\ \sigma_{12} & \sigma_2^2 & \sigma_{23}\\ \sigma_{13} & \sigma_{23} & \sigma_3^2\end{bmatrix}\right]\tag{16.47}$$

如同以往一样，只有 y_1^* 的符号是可观测时，才使用正规化 $\sigma_1^2=1$。

对数似然函数类似于 16.5 节二变量样本选择模型的情况，只是现在当 $y_1^*\leqslant 0$ 时，可观测到 y_3^*，因而式(16.33)中的项 $\Pr[y_{1i}^*\leqslant 0]$ 要用 $f(y_{3i}\mid y_{1i}^*\leqslant 0)\Pr[y_{1i}^*\leqslant 0]$ 代替。

估计模型更广泛使用的方法，是把赫克曼两步方法用于截尾均值：

$$\begin{aligned} \mathrm{E}[y|\mathbf{x}, y_1^* > 0] &= \mathbf{x}_2'\boldsymbol{\beta}_2 + \sigma_{12}\lambda(\mathbf{x}_1'\boldsymbol{\beta}_1) \\ \mathrm{E}[y|\mathbf{x}, y_1^* \leqslant 0] &= \mathbf{x}_3'\boldsymbol{\beta}_3 + \sigma_{13}\lambda(-\mathbf{x}_1'\boldsymbol{\beta}_1) \end{aligned} \quad (16.48)$$

其中，$\lambda(z)=\phi(z)/\Phi(z)$，而且我们使用 $\sigma_1^2=1$。不管怎样，$y_1^*>0$ 的第一阶段 probit 估计会得出 $\boldsymbol{\beta}_1$ 的估计值，从而得到 $\lambda(\mathbf{x}_1'\hat{\boldsymbol{\beta}}_1)$。于是，两个独立的 OLS 估计产生了$(\boldsymbol{\beta}_2, \sigma_{12})$与$(\boldsymbol{\beta}_3, \sigma_{13})$的估计值。然后，类似于式(16.40)后面二变量样本选择模型的技术，利用源自回归的残差平方，能够获得 σ_2^2 与 σ_3^2 的估计值。马达拉(Maddala, 1983，第 225 页)提供了该模型完整的详细内容，他称之为具有内生转换的转换回归模型(**switching regression model**)。这也是雨宫(Amemiya, 1985，第 399 页)曾经阐述的 Tobit 模型第五类。

16.7.2 *罗伊模型的变形*

许多模型可归入罗伊模型类。马达拉(Maddala, 1983，第 9 章)已经给出，他称之为具有自选择性模型的相关参考文献。也可参见雨宫(Amemiya, 1985，第 10 章)。此处，我们阐明几个重要例子。

二变量样本选择模型可被看成是下述特殊情况：忽略 y_3^* 且我们只对截尾矩 $\mathrm{E}[y_2^* | y_1^*>0]$加以建模。当 $y_1^* \leqslant 0$ 时，$y=0$ 的二变量样本选择模型，诸如在劳动力供给应用中，可更直接地看成是罗伊模型，其中我们要么观测到 $y=y_2^*$，要么观测到 $y=0$。因此，$y_3^*=0$。

在李龙飞(L. F. Lee, 1978)的研究中，y_2^* 与 y_3^* 分别表示工会工资与非工会工资，而 y_1^* 表示成为一个工会成员的意向。这增加了额外的结构：

$$y_1^* = y_2^* - y_3^* + \mathbf{z}'\boldsymbol{\gamma} + \zeta$$

其中，$\mathbf{z}'\boldsymbol{\gamma}+\zeta$ 反映出工会关系成本，同时更贴近罗伊(Roy, 1951)的思想。一旦代入 y_2^* 与 y_3^*，则得到 y_1^* 的简化式：

$$y_1^* = (\mathbf{x}_2'\boldsymbol{\beta}_2 - \mathbf{x}_3'\boldsymbol{\beta}_3 + \mathbf{z}'\boldsymbol{\gamma}) + (\varepsilon_2 - \varepsilon_3 + \zeta)$$

现在，这一模型与先前的模型相同，其修正项 $\lambda(\mathbf{x}_1'\hat{\boldsymbol{\beta}}_1)$可通过 y_1 对 $\mathbf{x}_1$ 的第一步回归来获得，其中，$\mathbf{x}_1$ 表示 $\mathbf{x}_2$、$\mathbf{x}_3$ 以及 $\mathbf{z}$ 中的唯一回归元。

若唯一截距由数量 α 表示，对于两个可能结果来说，它会变化，则罗伊模型简化成两个潜变量：

$$\begin{aligned} y_1^* &= \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1 \\ y^* &= \mathbf{x}_1'\boldsymbol{\beta}_1 + \alpha\, y_1 + \varepsilon_1 \end{aligned}$$

其中，$y=y^*$ 总是可观测到的，而且我们也可以观测到，当 $y_1^*>0$ 时，二值变量 y_1 等于 1，否则 y_1 等于 0。关于 y 的这个模型能被看成是具有虚拟内生变量(**dummy endogenous variable**)(y_1)的模型。它可通过应用赫克曼两步估计量的表达式 $\mathrm{E}[y^*|\mathbf{x}]$而得到估计。否则，倘若有关于 y_1 的工具可利用，则运用工具变量估计。这需要一种回归元，它不决定关注结果的水平，却决定哪个结果被选取。

这些罗伊模型类似于处理效应文献中所研究的模型。存在两个潜在结果，此处为 y_2^* 与 y_3^*，但我们仅仅能观测到两者之一。本章方法通过对不可观测因素的分布做强条件假设而建立了一种反事实框架。第 25 章将阐述其他一些可供选择的方法。特别地，参见 25.3 节关于各种不同方法之间的联系。

16.8 结构模型

关于选择样本的回归模型具有下述特性：关注结果部分地依赖于参与决策，而参与决策将反过来依赖于预期结果。参与决策与结果是同时决策。前面的表述借助于给出参与方程的简化形式(**reduced-form**)而简化这种相互依存性。特别地，参见 16.7.2 节中李(Lee，1978)的解释。这是一种有效方法，尽管与具有完全结构形式所起的作用相比稍欠有效。

在本节，我们利用建立在效用最大化基础上的结构经济模型，以显性方式对相互依存性加以建模，同时利用可将线性联立方程推广到包含删失与截尾的情况，包括二值结果的结构统计模型。

16.8.1 *基于效用最大化的结构模型*

最初，结构模型(**structural model**)研究考察女性劳动力供给。课本模型拥有消费者最大化效用的商品消费与闲暇时间函数，受限于预算约束与时间约束，在闲暇时间与工作时间之间进行配置的时间可自由决定如何利用。在内部解上，闲暇与商品消费之间的边际替代率(MRS)等于工资率。不过，若大于工资率，则会产生角点解，即妇女选择不参加工作。格罗诺(Gronau，1973)、赫克曼(Heckman，1974)都曾经阐述过与效用最大化相一致的经济计量模型，从而得到类似的 Tobit 模型，并解释了对于那些不参加工作的妇女来说，观测不到工资率的额外复杂情况。后来的研究包括并入工作的固定成本，产生了样本选择模型，并使用面板数据，从而产生面板模型。基林斯沃思和赫克曼(Killingsworth and Heckman，1986)、布伦德尔和麦柯迪(Blundell and MaCurdy，2001)提供了一个综述，而姆罗茨(Mroz，1987)则给出了一个应用。

为了阐明结构方法，我们概述下面的例子。迪宾和麦克法登(Dubin and McFadden，1984)将家庭的能源消费(电或天然气)与器具选择(比如电炉子或天然气炉子)建模成源自相同效用函数的相互联系的决策。特别地，假定 m 个器具组合中的第 j 个家庭间接效用(**indirect utility**)为：

$$V_j=\{\alpha_{0j}+\alpha_1/\beta+\alpha_1 p_1+\alpha_2 p_2+\mathbf{w}'\gamma+\beta(y-r_j)+\eta\}e^{-\beta p_1}+\varepsilon_j \qquad (16.49)$$

其中，p_1 与 p_2 分别表示电与天然气的价格，y 表示收入，而 r_j 表示按年度比率重估的组合的总管理生活周期成本，满足：

$$r_j=p_1q_{1j}+p_2q_{2j}+\rho c_j$$

其中，q_{1j} 与 q_{2j} 表示由拥有器具组合 j 的家庭所引起的典型电与气的消费，c_j 表示器具组合 j 的成本，而 ρ 表示贴现率。对于不同家庭，其喜好各不相同，这归因于

可观测的特性 $\mathbf{w}$、不可观测误差 η 以及器具组合特定误差 ϵ_j，它们被假定成关于 j 是不相关的，但关于 η 是相关的。此外，存在一个共同的器具喜好因子 α_{0j}。

给定器具组合 j 时，对电的需求等于 $-(\partial V_j/\partial p_1)/(\partial V_j/\partial y)$，由罗伊恒等式(**Roy's identity**)得到：

$$x_1-q_{1j}=\alpha_{0j}+\alpha_1 p_1+\alpha_2 p_2+\mathbf{w}'\gamma+\beta(y-r_j)+\eta$$

为了强调对器具组合 j 的选择是内生的，引入 m 个互不相交的指示变量 δ_{jk}，$k=1,\cdots,m$，其中：

$$\delta_{jk}=\begin{cases}1, & \text{当 } k=j\\ 0, & \text{当 } k\neq j\end{cases}$$

于是，给定器具组合 j 时对电的需求由下式给出：

$$x_1-q_{1j}=\sum_{k=1}^{m}\alpha_{0k}\delta_{jk}+\alpha_1 p_1+\alpha_2 p_2+\mathbf{w}'\gamma+\beta\Big(y-\sum_{k=1}^{m}r_j\delta_{jk}\Big)+\eta \tag{16.50}$$

即使模型(16.50)是线性的，OLS 回归也会由于 δ_{jk} 的内生性而产生非一致估计。迪宾和麦克法登(Dubin and McFadden, 1984)曾经阐述了其他两种可供选择的估计方法。

IV 方法(**IV approach**)利用 $\hat{p}_k$ 与 $r_j\hat{p}_k$ 作为 δ_{jk} 与 $r_j\delta_{jk}$ 的工具来估计式(16.50)，其中，$\hat{p}_k$ 表示选取各种不同器具组合的预测概率，$k=1,\cdots,m$。这里，V_j 用于表示间接效用函数。它既包括效用的确定成分，又包括效用的随机成分，同时对应于 15.5.1 节中表示式的 U_j。在 ϵ_j 是 iid 第 II 类极值，且 cdf $F(\epsilon)=\exp(-\exp(-\gamma-\epsilon\pi/\lambda\sqrt{3}))$ 的假设下，其中，$\gamma\simeq 0.577\,2$ 为欧拉常值，用类似方法得到：

$$\begin{aligned}p_k&=\Pr[V_k>V_l,\ l\neq k,\ l=1,\cdots,m]\\&=\Pr[\epsilon_l-\epsilon_k<\{(\alpha_{0k}-\alpha_{0l})-\beta(r_k-r_l)\}e^{-\beta p_1},\ \text{所有 } l\neq k]\\&=\frac{\exp[(\alpha_{0k}-\beta r_k)e^{-\beta p_1}\pi/\lambda\sqrt{3}]}{\sum_{l=1}^{m}\exp[(\alpha_{0l}-\beta r_l)e^{-\beta p_1}\pi/\lambda\sqrt{3}]}\end{aligned}$$

注意到，ϵ_j 具有零均值且方差为 $\lambda^2/2$，这些均不同于第 14 章与第 15 章所使用的第 II 类极值分布的那些参数化。对非线性多项式模型进行估计会得出预测概率 $\hat{p}_k$。

关于另一种可供选择的样本选择方法(**sample selection approach**)，注意到 $\mathrm{E}[\eta|$器具组合 $j]\neq 0$，同时使用 η 与 $\epsilon_1,\cdots,\epsilon_m$ 的分布假设来获得这个期望值。特别地，假定 $\eta|\epsilon_1,\cdots,\epsilon_m$ 是 iid 的，其均值为 $(\sqrt{2}\sigma/\lambda)\sum_{k=1}^{m}R_k\epsilon_k$，而方差为 $\sigma^2(1-\sum_{k=1}^{m}R_k^2)$，其中，$\sum_{k=1}^{m}R_k=0$ 且 $\sum_{k=1}^{m}R_k^2<1$，并且 ϵ_k 的分布已经给出的。然后，执行迪宾和麦克法登所给出的一些代数运算，得出[1]：

$$\mathrm{E}[\eta\mid \text{器具组合 } j]=\sum_{k\neq j}^{m}(\sigma\sqrt{6}R_k/\pi)\left[\frac{p_k\ln p_k}{1-p_k}+\ln p_j\right]$$

[1] 原著中这里最后一项表达式为 $\ln p_k$，但应为 $\ln p_j$。——译者注

于是，赫克曼两步法可通过 OLS 估计：

$$x_1 - q_{1j} = \sum_{k=1}^{m} \alpha_{0k}\delta_{jk} + \alpha_1 p_1 + \alpha_2 p_2 + \mathbf{w}'\gamma + \beta\left(y - \sum_{k=1}^{m} r_j\delta_{jk}\right) + \sum_{k\neq j}^{m}\gamma_k\left[\frac{\hat{p}_k \ln \hat{p}_k}{1-\hat{p}_k} + \ln \hat{p}_j\right] + \xi$$ [1]

其中，$\hat{p}_k$[2] 表示源自前面关于 p_k 模型的预测概率，而 ξ 表示具有渐近零均值的误差。

迪宾和麦克法登利用 3 249 户家庭使用两种可能供暖组合——电暖和气暖——的数据来估计这些模型。

有关例子包括哈恩曼（Hanemann，1984）对品牌消费水平的建模，其中，消费者在可能选择的品牌商品集中只能消费一种品牌，并且卡梅伦等人（Cameron et al.，1988）对在一系列互不相交的健康保险政策中选择其中一种的健康服务需求进行建模。

迪宾和麦克法登例子已表明，为了设定既对选择概率又对以选择为条件的需求进行解析的模型，需要许多创造力。甚至当不能获得解析解时。第 12 章与第 13 章所阐述的计算方法方面的进步允许对此类模型加以估计。不过，结果仍将依赖于所假定的效用以及不可观测因素的分布。

16.8.2 *联立方程 Tobit 模型与 probit 模型*

为了阐明推广 2.4 节线性方法涉及的问题，我们考察依赖于两个潜变量的选择模型，同时将联立性引进潜变量模型当中。一种相当一般的模型是：

$$\begin{aligned} y_1^* &= \alpha_1 y_2^* + \gamma_1 y_1 + \delta_1 y_2 + \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1 \\ y_2^* &= \alpha_2 y_1^* + \gamma_2 y_1 + \delta_2 y_2 + \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon_2 \end{aligned} \tag{16.51}$$

其中，y_1^* 与 y_2^* 均不是完全可观测的，却决定可观测变量 y_1 与 y_2，并假定误差服从联合正态分布。例如，当 $y_1^*>0$ 时，我们可观测到二值指示变量 $y_1=1$；而当 $y_2^*>0$ 时，可观测到 $y_2=y_2^*$。注意到，最重要的是，要么潜变量作为回归元，要么可观测结果作为回归元，或者这两者都可作为回归元而出现，尽管识别要求下面将给出的那些约束。

内生潜变量

最简单的是允许唯一潜变量成为式(16.51)的回归元。于是：

$$\begin{aligned} y_1^* &= \alpha_1 y_2^* + \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1 \\ y_2^* &= \alpha_2 y_1^* + \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon_2 \end{aligned} \tag{16.52}$$

二变量样本选择模型(16.31)是这样的例子，即另外设定 $\alpha_2=0$ 且直接设定 y_1^* 方程的简化式而不是结构式。很容易估计模型(16.52)，因为 y_1^* 与 y_2^* 的简化式能以通过与正规线性联立方程完全相同的方式来获得。然后，这一简化式可利用一

---

[1] 原著该公式中方括号内最后一项表达式为 $\ln \hat{p}_k$，但应为 $\ln \hat{p}_j$。——译者注

[2] 原著中这里为 p_k，疑似印刷错误，应为 $\hat{p}_k$。——译者注

些方法进行求解，比如利用依赖于给定 y_1^* 与 y_2^* 时，决定 y_1^* 与 y_2^* 方式的 probit 及 Tobit 方法。于是，结构模型（16.52）的参数，可通过运用简化式预测值 $\hat{y}_2^*$ 与 $\hat{y}_1^*$ 代替回归元 y_2^* 与 y_1^* 而得以估计。

把诸如式（16.52）的一些模型称为联立方程 Tobit 模型（**simultaneous equations Tobit models**）。如果可观测因变量 y_1 与 y_2 都是二值的，就产生了联立方程模型。纳尔逊和奥尔森（Nelson and OLson，1978）、雨宫（Amemiya，1979）以及李龙飞、马达拉和特罗斯特（Lee，Maddala，and Trost，1980）都提出这种估计量，而且李龙飞（Lee，1981）给出了一系列相当一般的研究。该估计量的标准误差，可利用 6.6 节关于序贯两步估计量的结果来获得。不过，更为简单的方法是利用 11.2 节阐述的成对自助法程序来获得。识别需要类似于那些线性联立方程的式（16.51）的排除性约束。

内生回归元

对模型（16.52）的一种普遍设定是具有内生回归元的模型，其中，内生回归元是完全可观测的。于是，y_2^* 是完全可观测的，因而，当 $y_1^*>0$ 时，我们观测到 $y_1=y_1^*$，否则 $y_1=0$。此模型变为：

$$\begin{aligned} y_1^* &= \alpha_1 y_2 + \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1 \\ y_2 &= \mathbf{x}'\boldsymbol{\pi} + v \end{aligned} \tag{16.53}$$

其中，第一个方程是关注的结构方程，而第二个方程是内生回归元 y_2 的简化式。再次注意到，这里 y_1 是连续的、非离散的。由于联合正态误差 $\varepsilon_1=\gamma v+\xi$，其中，$\xi$ 表示独立正态误差（参见 5.1 节），所以 $y_1^*=\alpha_1 y_2+\mathbf{x}_1'\boldsymbol{\beta}_1+\gamma v+\xi$。

两步估计方法，从 y_2 对 $\mathbf{x}$ 的回归中计算预测残差 $\hat{v}=y_2-\mathbf{x}'\hat{\boldsymbol{\pi}}$，然后从模型

$$y_1^* = \alpha_1 y_2 + \mathbf{x}_1'\boldsymbol{\beta}_1 + \gamma\hat{v} + e_1$$

中获得 Tobit 估计值，其中误差是正态分布。关于 y_2 的内生性检验能实施成利用源自 Tobit 软件包的标准误差的 $\gamma=0$ 沃尔德检验。该检验是线性模型中实施豪斯曼内生检验的辅助回归的推广（参见 8.4.3 节）。如果零假设被拒绝，那么前面提及的第二步回归得出 α_1 与 α_1 的一致估计值，可是标准误差则需要加以调整，其原因在于另外回归元 $\hat{v}$ 的第一步估计。关于 Tobit 模型的详细内容，参见史密斯和布伦德尔（Smith and Blundell，1986），而里弗斯和翁（Rivers and Vuong，1988）考察了第二步估计 probit 模型的类似方法。

内生删失或二值变量

如果式（16.51）中出现可观测的删失，或二值内生变量 y_1 或 y_2 作为回归元，那么分析就更为复杂。赫克曼（Heckman，1978）曾经考察下述模型：

$$\begin{aligned} y_1^* &= \gamma_1 y_1 + \delta_1 y_2^* + \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1 \\ y_2^* &= \alpha_2 y_1^* + \gamma_2 y_1 + \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon_2 \end{aligned} \tag{16.54}$$

其中，当 $y_1^*>0$ 时，我们观测到 $y_1=1$；当 $y_1^*\leqslant 0$ 时，观测到 $y_1=0$；而且在所有时间都可观测到 $y_2=y_2^*$。一种复杂情况是，此处 y_1 作为回归元出现。有意义的简化式只能依赖于 $\mathbf{x}_1$ 与 $\mathbf{x}_2$ 而不依赖于 y_1。这施加了 $\delta_1\gamma_2+\gamma_1=0$ 约束，即文献中所

谓的凝聚条件(coherence condition)的例子。于是,模型简化式变为:

$$y_1^* = \mathbf{x}'\boldsymbol{\pi}_1 + v_1$$
$$y_2 = \gamma_2 y_1 + \mathbf{x}'\boldsymbol{\pi}_2 + v_2$$

这是罗伊模型的一种特殊情况,其中参与($y_1=1$)导致唯一截距在结果中(经由 γ_2)移位。通常,对于具有包括删失或者截尾内生变量回归元的模型进行估计相当困难。例如,参见布伦德尔和史密斯(Blundell and Smith, 1989)。

例子

布鲁克斯、卡梅伦和卡特(Brooks, Cameron, and Carter, 1998)应用联立方程 Tobit 模型,解释议会代表对准糖修正案的投票。三个观测结果 y_1、y_2 以及 y_3 分别是投票(同意或反对)来自糖业利益集团对其竞争用专款的捐款以及对甜味剂使用的利益集团。第一个结果是二值结果,而其他两个结果都是在 0 点删失的。可设定有关的潜变量 y_1^*、y_2^* 以及 y_3^* 的联立方程模型,因而其结构模型具有较简单的式(16.52)的形式。

这个假定的合理性如何呢?竞争捐款 y_2^* 与 y_3^* 应该依赖于潜变量 y_1^*,因为真实投票 y_1 是在稍后日期做出的。然而,对于 y_1^*,一种可供选择的且更加困难的模型是,y_1^* 关于投票的潜变量 y_1^* 依赖于所接收的实际捐款(y_2 与 y_3)而不是潜在捐款。然而,如果这可以被认为可能是未来重复进行的博弈,那么该事情就利用 y_2^* 与 y_3^* 来完成。很明显,此类假设的合理性将随应用而变化。参数识别是通过关于外生回归元的排除性约束而得到保证。一致估计依赖于作为联合分布的误差。

16.9 半参数估计

删失、截尾以及样本选择均会导致不同于总体的样本。这在本质上是一个缺失数据问题,由于数据关于因变量而不是内生变量为缺失的,故这是一种复杂问题。前面一些方法,可通过做出分布假设解决此类缺失数据问题,要么获得样本数据的似然函数,要么获得适当删失、截尾或选择的条件均值。

这些方法甚至对于误差分布的极小错误设定来说都是脆弱的。例如,倘若误差是正态的且异方差的,或误差是同方差的且非正态的,则标准 Tobit 模型的 OLS 及其赫克曼两步估计量都是非一致的。例如,参见帕施(Paarsch, 1982)及其中的参考文献。

相当多的研究致力于发展半参数估计量,在较弱分布假设下,半参数估计量是一致的。可是,在阐述重要例子之前,我们提及一种可供选择的方法是,继续采用建立在更丰富、更灵活分布假设基础上的完全参数方法。

16.9.1 *灵活参数模型*

为了简单起见,以经典 Tobit 模型 $y_i^* = \mathbf{x}_i'\beta + \varepsilon_i$ 开始。其假设 $\varepsilon_i \sim \mathcal{N}[0, \sigma_i^2]$ 能以两种方式得以放松。首先,通过显性模型 $\sigma_i^2 = \exp(\mathbf{z}_i'\gamma)$ 来并入异方差性,这里

的 β 与 γ 是待估的。其次，可使用比正态分布更为灵活的分布。例如，人们能使用正态的平方多项式展开(参见 9.7.7 节)。

对于二变量样本选择模型，可采用类似方法，其中，现在使用($\varepsilon_1,\varepsilon_2$)更灵活的联合分布。为了使两变量正态性假设更为合理，李(Lee，1983)提出，以($\varepsilon_1,\varepsilon_2$)的变换($\varepsilon_1^*,\varepsilon_2^*$)进行研究。

贝叶斯方法也可用于此类模型。奇夫(Chib，1992)考察了删失模型。引进潜变量 $\mathbf{y}^*$ 作为辅助变量，并使用数据增广方法(参见 13.7 节)。吉布斯抽样器进行下述循环：(1) 关于 $\boldsymbol{\beta}\mid\mathbf{y},\mathbf{y}^*,\sigma^2$ 的条件后验；(2) 关于 $\sigma^2\mid\mathbf{y},\mathbf{y}^*,\boldsymbol{\beta}$ 的条件后验；(3) 关于 $\mathbf{y}^*\mid\mathbf{y},\boldsymbol{\beta},\sigma^2$ 的后验。

关于处理非线性模型中的删失、截尾以及样本选择，比如那些计数数据与持续期限数据或者混合的数据形式，当半参数方法很少能得以利用时，灵活参数方法(**flexible parametric approach**)尤其具有优势。

16.9.2 *删失模型的半参数估计*

现在，回到半参数估计上。我们考察潜变量的线性模型 $y_i^*=\mathbf{x}_i'\boldsymbol{\beta}+\varepsilon_i$，它是在 0 点左面删失的，因而当 $y_i^*>0$ 时，我们观测到 $y_i=y_i^*$，而当 $y_i^*\leqslant 0$ 时，则 $y_i=0$。半参数文献通常将这个模型表述成：

$$y_i=\max(\mathbf{x}_i'\boldsymbol{\beta}+\varepsilon_i,0) \tag{16.55}$$

这是 Tobit 模型(16.11)～(16.13)，只是 ε 的分布是未设定的。对该模型做某种改动，同样可涵盖在已知固定点而不是 0 点的左删失，以及右删失诸如上端编码数据的情况。例如，如果 $y=\min\{\mathbf{x}'\boldsymbol{\beta}+\varepsilon,U\}$，那么 $U-y=\max\{U-\mathbf{x}'\boldsymbol{\beta}-\varepsilon,0\}$。其目标是在没有设定 ε_i 的完整参数分布下，一致地估计 $\boldsymbol{\beta}$。此估计量称为半参数的，因为未删失均值 $\mathbf{x}_i'\boldsymbol{\beta}$ 是参数化的，而误差分布则不是参数化的。下面要阐述的方法在对 ε 分布所做出的假设上存在差异。

由式(16.8)知，给定 y^* 的 cdf 知识，进而是 ε 的知识，进行 ML 估计是可行的。对右删失持续期限数据情况，利用第 17 章阐述的关于 cdf 卡普兰-麦耶(Kaplan-Meier)乘积极限估计量能以非参数形式估计。否则，ε 的分布可利用加伦特和尼奇卡(Gallant and Nychka，1987)的序列展开以非参数形式得以确定，参见 9.7.7 节，这些半参数 ML 估计方法极少得到应用。

然而，文献关注于基于条件矩的估计。由式(16.20)知，条件删失均值 $\mathrm{E}[y\mid\mathbf{x}]$ 显然是单指标模型，满足 $\mathrm{E}[y\mid\mathbf{x}]=g(\mathbf{x}'\boldsymbol{\beta})$，其中，若对 ε 分布不加以设定，则函数 $g(\cdot)$ 是未知的。因此，9.7.4 节的单指标方法就能得到应用，尽管如同注意到的，对 $\boldsymbol{\beta}$ 加以估计，至多仅相差一个位置与标度。

一种更流行的方法考察可供选择的条件删失矩，该矩很少受到删失的改变。鲍威尔(Powell，1984)提出利用条件中位数(**conditional median**)。重要的分布假设是，$\varepsilon\mid\mathbf{x}$ 具有零中位数，在此情况下，$y\mid\mathbf{x}$ 的条件中位数等于条件均值 $\mathbf{x}'\boldsymbol{\beta}$。通过假定 y 是 iid 的，最容易获得鲍威尔估计量的一种直觉认知。如果删失是对少于样本一半进行的，以至于少于观测值一半的为 0，且多于观测值一半的为正的，那么

删失样本增位数提供了总体位数的一致估计。鲍威尔(Powell, 1984)将这种思想推广到回归情况,对于少于观测值一半的 $\varepsilon|\mathbf{x}$ 进行删失,就上述那些观测值实施同样逻辑,其中,$\varepsilon=y-\mathbf{x}'\boldsymbol{\beta}$依赖于$\boldsymbol{\beta}$,而$\boldsymbol{\beta}$是需要估计的。中位数估计的回归类似形式是 LAD 估计(参见 4.6 节)。这就产生了删失最小绝对偏差(CLAD)估计量(**censored least absolute deviations estimator**)$\hat{\boldsymbol{\beta}}_{\text{CLAD}}$,它极小化:

$$Q_N(\boldsymbol{\beta}) = N^{-1}\sum_{i=1}^{N}|y_i-\max(\mathbf{x}_i'\boldsymbol{\beta},0)| \tag{16.56}$$

此估计量一致性的根本条件是,$\varepsilon|\mathbf{x}$ 具有零中位数。给定这一假设,此估计量是$\sqrt{N}$一致的,即使误差是条件异方差的。$\boldsymbol{\beta}$ 的估计量是一致的且渐近正态的。更有效估计量能够通过以 $f(0|\mathbf{x}_i)$对和式中的一些项进行加权获得,$\varepsilon_i|\mathbf{x}_i$ 的条件密度在 0 点进行计算。该方法同样可以被推广到条件分位数上。

一种可供选择的方法是,使用对称调整均值(**symmetrically trimmed mean**)而不是均值,这也不会受到删失的影响。假定 $\varepsilon|\mathbf{x}$ 分布是对称分布。这蕴含,对于具有正均值的观测值(即 $\mathbf{x}'\boldsymbol{\beta}>0$),在$(0,\ 2\mathbf{x}'\boldsymbol{\beta})$上服从对称分布。于是,或者$\mathbf{x}'\boldsymbol{\beta}+\varepsilon<0$ 与 $y=0$ 是可观测到的,或者以相等概率 $\mathbf{x}'\boldsymbol{\beta}+\varepsilon>2\mathbf{x}'\boldsymbol{\beta}$,同时为了确保关于 $\mathbf{x}'\boldsymbol{\beta}$对称,而人为地将数据设置成 $2\mathbf{x}'\boldsymbol{\beta}$。

我们已经证明:

$$\mathrm{E}[\mathbf{1}(\mathbf{x}'\boldsymbol{\beta}>0)[\min(y,2\mathbf{x}'\boldsymbol{\beta})-\mathbf{x}'\boldsymbol{\beta}]\mathbf{x}]=\mathbf{0} \tag{16.57}$$

其中,$\mathbf{1}(\mathbf{x}'\boldsymbol{\beta}>0)$把注意力限制在具有正均值上,而且新的因变量是 $y=0$,或者$0<y<2\mathbf{x}'\boldsymbol{\beta}$,或者若 $y>2\mathbf{x}'\boldsymbol{\beta}$,则为 $2\mathbf{x}'\boldsymbol{\beta}$。建立在式(16.57)基础上的矩估计量没有 $\boldsymbol{\beta}$的唯一解。鲍威尔(Powell, 1986b)提出了对称删失最小二乘法[**symmetrically censored least squares (SCLS) estimator**]估计量,该估计量极小化:

$$Q_N(\boldsymbol{\beta})=N^{-1}\sum_{i=1}^{N}\{[y_i-\max(y_i/2,\mathbf{x}_i'\boldsymbol{\beta})]^2+\mathbf{1}(y_i>2\mathbf{x}_i'\boldsymbol{\beta})[y_i^2/4-\max(0,\mathbf{x}_i'\boldsymbol{\beta})]^2\} \tag{16.58}$$

经过一些代数运算,可以证明,得到的一阶条件是矩条件(16.57)的样本类似形式。斋和奥诺雷(Chay and Honoré, 1998)提供了 SCLS 估计量修饰的图形解释,并且奥诺雷和鲍威尔(Honoré and Powell, 1994)提供了相对分段差分估计量。

梅伦伯格和范·索斯特(Melenberg and Van Soest, 1996),斋和奥诺雷(Chay and Honoré, 1998)以及斋和鲍威尔(Chay and Powell, 2001)都曾经给出这些估计量中某些估计量的应用。帕甘和乌拉(Pagan and Ullah, 1999)提供另外一些方法及理论。

举一个实证例子,我们将 CLAD 估计应用于 16.2.1 节的数据,这是由具有正态误差的 Tobit 模型生成的数据。利用 ML 估计的斜率参数(设定为 1 000)是 956(标准误差 117),与利用 CLAD 所得到的斜率参数 838(标准误差 165)相比较。正如人们所料,CLAD 对非正态性的稳健是以有效性的某种损失为代价的。

16.9.3 *样本模型的半参数估计*

对样本选择模型进行半参数估计是更富有挑战性的。我们考察最广泛研究的

模型，即 16.5.2 节定义的二变量样本选择模型(**bivariate sample selection model**)，其中，现在将误差(ε_1，ε_2)服从联合正态分布的假设加以放松。

半参数 ML 估计是可行的。特别地，加伦特和尼奇卡(Gallant and Nychka，1987)以显性方式，考察了二变量样本选择模型作为 9.7.7 节曾阐述的级数展开估计量的合适备选者。

不过，文献却以截尾条件均值表达式作为起点，由式(16.3.4)知，截尾条件均值由

$$\begin{aligned}\mathrm{E}[y_{2i}|\mathbf{x}_i, y_{1i}^*>0] &= \mathbf{x}_{2i}'\boldsymbol{\beta}_2+\mathrm{E}[\varepsilon_2|\varepsilon_1>-\mathbf{x}_{1i}'\boldsymbol{\beta}_1] \\ &= \mathbf{x}_{2i}'\boldsymbol{\beta}_2+g(\mathbf{x}_{1i}'\boldsymbol{\beta}_1)\end{aligned} \tag{16.59}$$

给出，其中第二个等式假定：$\varepsilon_{2i}|\mathbf{x}_i$，$\varepsilon_{1i}$ 具有只依赖于 $\mathbf{x}_{1i}$ 的分布，这类似于式(16.41)。(ε_1，ε_2)的分布是未设定的，因而函数 $g(\cdot)$ 是未知的，从而出现半参数估计问题。由于 $g(\mathbf{x}_1'\boldsymbol{\beta}_1)=(\mathbf{x}_1'\boldsymbol{\beta}_1)$是可能的，识别具有未设定的该模型就需要下面这个排除性约束：$\mathbf{x}_1$ 中的至少一个分量不出现在 $\mathbf{x}_2$ 中。更进一步地，$\mathbf{x}_1'\boldsymbol{\beta}_1$ 与 $\mathbf{x}_2$ 越是不相关，则 $\boldsymbol{\beta}_2$ 与 $g(\cdot)$ 就越能更好地加以区分。模型(16.59)是偏线性模型(部分线性模型)，它可利用 9.7.3 节阐述的方法进行估计。一些流行的方法包括鲁宾逊(Robinson，1988a)的差分估计量以及利用的级数展开。由于 $\boldsymbol{\beta}_1$ 是未知的，y_{2i}对 $\mathbf{x}_{2i}'\boldsymbol{\beta}_2+g(\mathbf{x}_{1i}'\hat{\boldsymbol{\beta}}_1)$进行回归，利用 14.7 节给出的半参数二值模型估计量之一加以估计，其中，$\hat{\boldsymbol{\beta}}_1$ 能通过 y_{1i}对 $\mathbf{x}_{1i}$的二值结果回归获得。这些方法提供了斜率参数 $\boldsymbol{\beta}_2$ 的一致估计。为了另外估计截距，必须对原水平值而不是 y_2 变化进行分析，参见安德鲁斯和谢夫根斯(Andrews and Schafgens，1998)。

纽韦、鲍威尔和沃克(Newey，Powell，and Walker，1990)将此方法用于女生劳动力供给上。参与指示变量模型可利用几种不同方法估计，并且结果 y_2 的方程可利用鲁宾逊(Robinson，1988a)的方法加以估计。梅伦伯格和范・索斯特(Melenberg and Van Soest，1996)对假期支出利用一系列广泛的半参数方法进行建模，其中，既有二变量样本选择模型，又有删失回归模型。达斯、纽韦和维拉(Das，Newey and Vella，2003)提供了较丰富的模型。

曼斯基(Manski，1989)在相对最小假设下，考察了二变量样本选择模型的识别，并给出既以回归元为条件又以选择为条件的均值界与边际效应界。

16.10 推导 Tobit 模型

16.10.1 *标准正态的截尾矩*

考察 $z\sim\mathcal{N}[0,1]$，具有密度 $\phi(z)=(1/\sqrt{2\pi})\exp(-z^2/2)$和 cdf $\Phi(z)$。由于 $\Pr[z>c]=1-\Phi(c)$，所以 $z|z>c$ 的条件密度是 $\phi(z)/(1-\Phi(z))$。由此可得：

$$\begin{aligned}\mathrm{E}[z|z>c] &= \int_c^\infty z(\phi(z)/[1-\Phi(c)])dz \\ &= \int_c^\infty z(1/\sqrt{2\pi})\exp(-z^2/2)dz\Big/[1-\Phi(c)]\end{aligned}$$

$$= \int_c^{\infty} \frac{\partial}{\partial z}(-(1/\sqrt{2\pi})\exp(-z^2/2))dz \Big/ [1-\Phi(c)]$$

$$= [-(1/\sqrt{2\pi})\exp(-z^2/2)]_c^{\infty} \Big/ [1-\Phi(c)]$$

$$= \phi(c)/[1-\Phi(c))]$$

类似地，有：

$$\mathrm{E}[z^2 \mid z > c] = \int_c^{\infty} z^2(\phi(z)/[1-\Phi(c)])dz$$

$$= \int_c^{\infty} z \times z \times (1/\sqrt{2\pi})\exp(-z^2/2)dz \Big/ [1-\Phi(c)]$$

$$= \int_c^{\infty} z \times \frac{\partial}{\partial z}(-(1/\sqrt{2\pi})\exp(-z^2/2))dz \Big/ [1-\Phi(c)]$$

$$= [z \times (-1/\sqrt{2\pi})\exp(-z^2/2)]_c^{\infty} \Big/ [1-\Phi(c)]$$

$$- \int_c^{\infty} \frac{\partial}{\partial z}(z) \times (-(1/\sqrt{2\pi})\exp(-z^2/2))dz \Big/ [1-\Phi(c)]$$

$$= c\phi(c)/[1-\Phi(c)] + (1-\Phi(c))/[1-\Phi(c)]$$

$$= c\phi(c)/[1-\Phi(c)] + 1$$

经过一些代数运算之后，可得：

$$\mathrm{V}[z \mid z > c] = \mathrm{E}[z^2 \mid z > c] - (\mathrm{E}[z \mid z > c])^2$$

$$= 1 + c\phi(c)/[1-\Phi(c)] - \phi(c)^2/[1-\Phi(c)]^2$$

16.10.2 *Tobit 模型赫克曼两步估计量的渐近理论*

由于两步赫克曼估计量依赖于第一步参数估计值，所以该估计量的渐近方差矩阵极为复杂。存在几种方法获得渐近方差，比如雨宫的方法（Amemiya，1985，第 369～370 页）。然而，这里我们应用由 6.6 节给出的序贯两步估计量的一般结果。考察 Tobit 模型的最简单估计量（参见 16.3.6 节）。一些方法适合于二变量样本选择模型（16.5.4 节）以及联立方程 Tobit 模型（16.8.2 节）的两步估计量。一种更为简单的截然不同的方法是，使用自助法成对方法（参见 11.2 节）。

由式（16.26）知，我们希望估计关于正 y_i 的方程：

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \sigma\lambda(\mathbf{x}_i'\boldsymbol{\alpha}) + \eta_i$$

$$= \mathbf{w}_i(\boldsymbol{\alpha})'\boldsymbol{\gamma} + \eta_i$$

中的参数 $\boldsymbol{\gamma} = [\boldsymbol{\beta}' \quad \sigma]'$，其中，$\mathbf{w}_i(\boldsymbol{\alpha}) = [\mathbf{x}_i' \quad \lambda_i(\mathbf{x}_i'\boldsymbol{\alpha})]'$，而 $\eta_i = y_i - \mathbf{x}_i'\boldsymbol{\beta} - \sigma\lambda(\mathbf{x}_i'\boldsymbol{\alpha})$ 表示具有式（16.24）所定义的方差的异方差性。两步方法的第一步是要通过probit ML 获得未知参数 $\boldsymbol{\alpha}$ 的估计值 $\hat{\boldsymbol{\alpha}}$。由此可得，赫克曼两步估计量的两个部分正规方程为：

$$\sum_{i=1}^{N}(y_i - \Phi(\mathbf{x}_i'\boldsymbol{\alpha}))\frac{\phi^2(\mathbf{x}_i'\boldsymbol{\alpha})}{\Phi(\mathbf{x}_i'\boldsymbol{\alpha})(1-\Phi(\mathbf{x}_i'\boldsymbol{\alpha}))}\mathbf{x}_i = \mathbf{0} \qquad (16.60)$$

$$-\sum_{i=1}^{N} d_i\mathbf{w}_i(\boldsymbol{\alpha})(y_i - \mathbf{w}_i(\boldsymbol{\alpha})'\boldsymbol{\gamma}) = \mathbf{0}$$

其中，第一个方程给出了 $\boldsymbol{\alpha}$ 的 probit 一阶条件，而第二个方程给出了 $\boldsymbol{\gamma}$ 的关于正的 $y_i(d_i=1)$ 的 OLS 一阶条件。

这些方程能组合成 $\sum_{i=1}^{N}\mathbf{h}(\mathbf{x}_i,\boldsymbol{\theta})=\mathbf{0}$，其中，$\boldsymbol{\theta}=(\boldsymbol{\alpha}',\boldsymbol{\gamma}')'$。利用通常一阶泰勒级数展开，$\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\xrightarrow{d}\mathcal{N}[\mathbf{0},\ \mathbf{G}_0^{-1}\mathbf{S}_0(\mathbf{G}_0^{-1})']$，其中，$\mathbf{G}_0=\lim N^{-1}\mathrm{E}[\sum_{i=1}^{N}\partial\mathbf{h}(\mathbf{x}_i,\boldsymbol{\theta})/\partial\boldsymbol{\theta}]$，而 $\mathbf{S}_0=\lim N^{-1}\mathrm{E}[\sum_{i=1}^{N}\partial\mathbf{h}(\mathbf{x}_i,\boldsymbol{\theta})\mathbf{h}(\mathbf{x}_i,\boldsymbol{\theta})]'$。我们对相应于 $\boldsymbol{\gamma}$ 的子分量感兴趣。因为 $\partial\mathbf{h}(\mathbf{x}_i,\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ 是分块三角的，由于 $\boldsymbol{\gamma}$ 没有出现在第一个方程组中，所以可出现简化。一旦分割处理，得到一般结果：

$$\mathrm{V}[\hat{\boldsymbol{\theta}}_2]=\mathbf{G}_{22}^{-1}\{\mathbf{S}_{22}+\mathbf{G}_{21}[\mathbf{G}_{11}^{-1}\mathbf{S}_{11}\mathbf{G}_{11}^{-1}]\mathbf{G}_{21}'-\mathbf{G}_{21}\mathbf{G}_{11}^{-1}\mathbf{S}_{12}-\mathbf{S}_{21}\mathbf{G}_{11}^{-1}\mathbf{G}_{21}'\}\mathbf{G}_{22}^{-1}$$

其中，矩阵已在 6.6 节定义过。

若对这里的问题进行专门研究，我们首先考察 $\mathbf{G}_0$ 中的一些项。于是：

$$\mathbf{G}_{11}=\lim\frac{1}{N}\sum_{i=1}^{N}\frac{\phi^2(\mathbf{x}_i'\boldsymbol{\alpha})}{\Phi(\mathbf{x}_i'\boldsymbol{\alpha})(1-\Phi(\mathbf{x}_i'\boldsymbol{\alpha}))}\mathbf{x}_i\mathbf{x}_i'$$

$$\mathbf{G}_{21}=\lim\frac{1}{N}\sum_{i=1}^{N}d_i\mathbf{w}_i\frac{\partial\boldsymbol{\lambda}(\mathbf{x}_i'\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}}$$

$$\mathbf{G}_{22}=\lim\frac{1}{N}\sum_{i=1}^{N}\mathrm{E}[d_i\mathbf{w}_i\mathbf{w}_i']$$

$\mathbf{G}_{11}$ 的表达式使用了刚好 $\mathbf{G}_{11}^{-1}$ 是 probit MLE 的方差知识。$\mathbf{G}_{21}$ 的表达式使用了：

$$\begin{aligned}\mathrm{E}\left[\frac{\partial\mathbf{h}_{2i}}{\partial\boldsymbol{\theta}_1'}\right]&=\mathrm{E}\left[-\frac{\partial d_i\mathbf{w}_i(\boldsymbol{\alpha})(y_i-\mathbf{w}_i(\boldsymbol{\alpha})'\boldsymbol{\gamma}}{\partial\boldsymbol{\alpha}}\right]\\&=\mathrm{E}\left[\mathbf{w}_i\frac{\partial d_i\mathbf{w}_i(\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}'}\right]\\&=\mathrm{E}\left[d_i\mathbf{w}_i\frac{\partial\boldsymbol{\lambda}(\mathbf{x}_i'\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}}\right]\end{aligned}$$

$\mathbf{G}_{22}$ 的表达式使用了：

$$\frac{\partial\mathbf{h}_{2i}}{\partial\boldsymbol{\theta}_2'}=\frac{\partial d_i\mathbf{w}_i(\boldsymbol{\alpha})(y_i-\mathbf{w}_i(\boldsymbol{\alpha})'\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}=d_i\mathbf{w}_i\mathbf{w}_i'$$

转回到 $\mathbf{S}_0$ 上，我们有：

$$\mathbf{S}_{11}=\mathbf{G}_{11}^{-1}$$

$$\mathbf{S}_{21}=\mathbf{0}$$

$$\mathbf{S}_{22}=\lim\frac{1}{N}\sum_{i=1}^{N}\mathrm{E}[d_i(y_i-\mathbf{w}_i(\boldsymbol{\alpha})'\boldsymbol{\gamma})^2]$$

通过利用信息矩阵等式，可得到 $\mathbf{S}_{11}$ 的表达式。取数学期望且经过某些运算，得到 $\mathbf{S}_{21}=\mathbf{0}$，而且 $\mathbf{S}_{22}$ 正是 $\mathrm{V}[\eta_i]$。

对这些结果加以整理组合，得到赫克曼两步估计量 $\hat{\boldsymbol{\gamma}}\overset{a}{\sim}\mathcal{N}[\boldsymbol{\gamma},\ \mathbf{V}_\gamma]$，其中：

$$\hat{\mathbf{V}}_\gamma=(\hat{\mathbf{W}}'\hat{\mathbf{W}})^{-1}(\hat{\mathbf{W}}'\boldsymbol{\Sigma}_{\hat{\eta}}\hat{\mathbf{W}}+\hat{\mathbf{W}}'\hat{\mathbf{D}}\hat{\mathbf{V}}_\alpha\hat{\mathbf{D}}\hat{\mathbf{W}})(\hat{\mathbf{W}}'\hat{\mathbf{W}})^{-1}\tag{16.61}$$

而且 $\hat{\mathbf{W}}'\hat{\mathbf{W}}=\sum_{i=1}^{N}d_i\hat{\mathbf{w}}_i\hat{\mathbf{w}}_i'$，$\hat{\mathbf{D}}=\mathrm{Diag}[\partial\lambda(\mathbf{x}_i'\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}\mid_{\hat{\alpha}}]$，$\hat{\mathbf{V}}_{\boldsymbol{\alpha}}$ 表示第一阶段 probit MLE

的方差矩阵，同时 $\mathbf{\Sigma}_{\hat{\eta}}$ 表示第 i 个元素为 $\hat{\sigma}^2_{\eta_i}$ 的对角矩阵。若利用矩阵命令，则直接获得该估计值。最困难部分是以解析形式获得式(16.24)给出的 $\hat{\sigma}^2_{\eta_i}=\mathrm{V}[\eta_i]$。然后，假如这样做很困难，我们反而遵循怀特(White, 1980)的方法，使用 $\hat{\sigma}^2_i=(y_i-\mathbf{x}_i'\hat{\boldsymbol{\beta}}+\hat{\sigma}\lambda_i(\mathbf{x}_i'\hat{\boldsymbol{\alpha}}))^2$。

16.11 应用研究

绝大多数重要软件都包括正态条件下 Tobit 模型的 ML 估计。由于人们可分别估计两部分模型的每一个部分，所以很容易估计两部分模型。原则上，二变量样本选择模型，可借助于仅仅利用 probit 与 OLS 方法的赫克曼两步方法得到估计。然而，由于估计量的两步特性，很难计算其标准误差，可利用具有嵌入的赫克曼两步方法软件包，更容易获得标准误差。实施半参数估计量需要运用诸如 GAUSS 程序语言进行专门编程。一些软件包也会允许执行其他模型的删失变形与截尾变形的 ML 估计，比如计数数据的泊松或负二项模型。

假如人们将删失与截尾看成是合理的特定分布，就容易处理它们。例如，若对数正态分布拟合数据表现得很好，则容易处理上端编码的收入数据。删失 LAD 依赖于更弱的分布假设，故删失 LAD 也能用于此类情况中。

更为严重的问题是处理带有样本选择的模型。这些模型的更多参数形式均依赖于如下分布假设，这种分布假设使人认为是一种强假设。半参数形式仍必须努力满足识别要求，即决定参与的变量不能决定关注结果。一种更有前途的方法，也是人们在处理效应文献中经常采取的途径，是将注意力限制在如下情况：有理由假定，选择仅是依据可观测因素做出的。

16.12 文献注释

源自选择样本的有关模型文献浩如烟海。具有教科书篇幅的见解则由马达拉(Maddala, 1983)与古里耶克斯(Gouriéroux, 2000)做出，而较简短的概述由雨宫(Amemiya, 1984,1985)以及格林(Greene, 2003)给出。

16.3 托宾(Tobit, 1958)提出 Tobit 模型，并将它应用于消费支出数据上。雨宫(Amemiya, 1973)正式建立起该模型的一致性与渐近正态性。赫克曼(Heckman, 1974)提供了女性劳动力供给的出色应用，并详细分析了结果。

16.4 许多对兰德健康保险实验的研究，譬如端等人(Duan et al., 1983)，都是两部分模型的重要应用。

16.5 赫克曼(Heckman, 1976,1979)阐述了二变量样本选择模型的两步估计量，这也是最近众多半参数估计方法的基础。姆罗茨(Mroz, 1987)给出了一个女性劳动力供给的优秀应用，他强调了工资外生性假设的作用。

16.7 正如 Tobit 模型存在许多变形一样，罗伊思想也有众多变化形式。李龙飞(L. F. Lee, 1978)提供了早期对工会--非工会工资差异的应用。

16.8 由迪宾和麦克法登(Dubin and McFadden, 1984)做出的一项研究工

作，是结构微观经济计量分析的重要例子，该分析建立在对效用函数与不可观测因素的分布完全设定基础上。

16.9 二值选择模型的半参数估计已由李明宰（M-J. Lee，1996）、赫尔维茨（Horowitz，1997）以及帕甘和乌拉（Pagan and Ullah，1999）的书详细阐述，而综述是由维拉（Vella，1998）和李龙飞（L. F. Lee，2001）给出的。斋和奥诺雷（Chay and Honoré，1998）以及斋和鲍威尔（Chay and Powell，2001）都给出了删失模型的一些应用，另外，梅伦伯格和范·索斯特（Melenberg and Van Soest，1996）估计了二变量样本选择模型。

习 题

16-1 本题考察 Tobit 模型中各种不同截尾的影响。

(a) 生成潜变量 $y^*=k+3x+u$ 的 200 个采样，其中 $u\sim\mathcal{N}[0,3]$，回归元 $x\sim\mathcal{U}[0,1]$。选择 k，使得你生成的 y^* 大致有 30%成为负的。

(b) 通过排除对应于 $y^*<0$ 的观测值，生成删失或截尾子样本。

(c) 利用 200[1] 个观测值，通过 OLS 估计此模型，就好像潜变量是可观测的一样。

(d) 仅仅利用 $y>0$ 的截尾子样本，通过 OLS 估计此模型。

(e) 利用所有观测值，使用截尾极大似然选项估计参数。依照截尾 MLE 的性质，评价你的结果。将最小二乘法结果与前面两个部分的结果加以比较。

(f) 为了生成 20%、40%以及 50%的删失观测值，请利用 k 值，重复上面所有步骤。由此，对于较高水平的删失的参数估计，你会提出什么结果呢？如有可能，请利用理论强化你的推断。

16-2 考察由 $y_i^*=\mathbf{x}_i'\boldsymbol{\beta}+\varepsilon_i$ 建立的潜变量模型，其中，$\varepsilon_i\sim\mathcal{N}[0,\sigma^2]$。假定 y_i^* 从上面删失，因此，当 $y_i^*<U_i$ 时，我们观测到 $y_i=y_i^*$，当 $y_i^*\geqslant U_i$，观测到 $y_i=U_i$，其中，上限 U_i 表示对每个个体而言都为已知常值（也就是数据），但对所有个体而言可能是变化的。

(a) 给出这个模型的对数似然函数。（提示：注意到，不同于标准情况，因为 U_i 存在，同时若 $y_i^*<U_i$，则等式反过来满足 $y_i=y_i^*$。）

(b) 求出截尾均值表达式 $\mathrm{E}[y_i|\mathbf{x}_i,\ y_i<U_i]$。（提示：对于 $z\sim\mathcal{N}[0,1]$，我们有 $\mathrm{E}[z|z>c]=\phi(c)/[1-\Phi(c)]$。同样地，$\mathrm{E}[z|z<c]=-\mathrm{E}[-z|-z>-c]$，而 $-z\sim\mathcal{N}[0,1]$。）

(c) 由此，给出该模型的赫克曼两步估计量。

(d) 求出删失均值表达式 $\mathrm{E}[y_i|\mathbf{x}_i]$。[提示：(b)部分的解答是基础部分。]

16-3 此问题考察 Tobit 模型错误设定的后果。起点是习题 16.1 模型。

(a) 通过令 $u\sim\mathcal{N}[0,\sigma^2 z]$，估计具有异方差性的 y^*，其中，$z>0$ 表示选取合适的正值变量与 x 相关，尽管不完全如此。再一次选择 k，获得大致 30%的删失观

[1] 原著中这里为 2 000，但应该为 200。——译者注

测值。利用正常状态删失的 MLE 来估计该模型,同时将你的结果与相应的同方差情况进行比较。

(b) 现在考察样本非正态性的影响。使用某些软件包中可利用的非常大的模拟,完成基于 1 000 个观测值样本和 500 次复制的蒙特卡罗估值。在每一次复制中,生成具有删失观测值的样本,使得误差从两个正态分布即 $\mathcal{N}[1, 9]$ 或 $\mathcal{N}[0.4, 1]$的混合中采样,即分别以概率 0.4 与概率 0.6 进行采样。利用删失 Tobit MLE 估计该模型,并将你的结果与正态情况进行比较。完成对两个估计量的蒙特卡罗输出分析。推导 Tobit 估计量的非正态性分布影响的适当结论。

16-4 考察泊松回归模型,其中,y^* 具有密度 $f^*(y^*)=e^{-\mu}\mu^{y}/y^*!$, $y_i^*=0, 1, 2, \cdots$,并且对于不同 i 具有独立性。由于编码误差的缘故,我们只有当 $y^*\geqslant 2$ 时,才能完全观测到 y^*,当 $y^*=0$ 或 $y^*=1$ 时,才能观测到 $y^*\leqslant 1$。假定当 $y^*=1$ 时,可以编码。对于 $y_i^*\geqslant 2$,定义观测数据 $y=y^*$,而对于 $y_i^*=0$ 或 $y_i^*=1$,定义 $y=1$。

(a) 求出观测到 y 的密度 $f(y)$。

(b) 求出 $\mathrm{E}[y]$。(这里,要经过某些代数运算。)

现在引进满足 $\mathrm{E}[y^*|\mathbf{x}]=\exp(\mathbf{x}'\boldsymbol{\beta})$的回归元,并且对于 $y^*\geqslant 2$,定义指示变量 $d=1$,而对于 $y^*=0$ 或 $y^*=1$,定义 $d=0$。

(c) 给出这个例子中估计量目标函数的准确表达式,该估计量利用关于 y_i、d_i 以及 $\mathbf{x}_i$ 的数据,提供 $\boldsymbol{\beta}$ 的一致估计量。

(d) 给出这个例子中估计量目标函数的准确表达式,该估计量仅仅利用关于 d_i 以及 $\mathbf{x}_i$ 的数据,提供 $\boldsymbol{\beta}$ 的一致估计量。

(e) 仅利用关于 d_i 与 $\mathbf{x}_i$ 的数据,可能一致估计出 $\boldsymbol{\beta}$ 吗? 请解释你的回答。

16-5 利用本章的全部 12 个月的医疗消费支出数据的 50%随机子样本,并利用类似模型设定,我们希望考察下面的广泛问题,就消费支出数据建模而言,哪一种模型是适宜的?

(a) 利用消费支出变量的数据概括统计量,分析观测到 0 消费支出的较高比例的含义。这是否违背了正态性假设? 存在消费支出的那种变换吗? 该变换会促使所做出的正态性假设更为适合。

(b) 考察三个备选模型,每一个都具有相同的协变量集合。这些协变量与计数数据的习题 20-6 相同。这三个模型分别是:(1) Tobit 模型;(2) 两部分(围栏)模型(TPM);(3) 选择模型。请解释如何建立这些模型的每一种,它们之间的联系如何,并且人们如何去比较并选择它们。假如你遇到任何一种特定的设定或者估计问题,就阐述它们,并提出你是如何解决它们的。注意排除性约束的选择。

(c) 依次估计 Tobit 模型、TPM 以及选择模型。对于 TPM,你有两个方程,第二个方程是针对仅有正消费支出的那些人。就选择模型而言,运用 MLE 估计量与(赫克曼)两步估计量。讨论你估计选择模型时所需的排除性约束的理论。确实存在证据表明选择问题是一个严重问题吗?

(d) 如何比较这三种模型的拟合数据情况。哪种模型看起来提供了最佳数据拟合? 凭借什么准则来评判?

(e) 假如主要关注内容在于：两个变量即收入对数、(1＋共保率)对数对支出的影响。运用你估计 Tobit 与 TPM 的结果，比较这两种变量变动对支出产生的边际效应。倘若样本存在相当大的异质性，你怎样以最有信息价值的方式阐述分析结果？

(f) 简略解释分位数回归(参见 4.6 节)提供分析同样数据的另一种方法，针对目前数据，该方法的主要优缺点分别是什么？

17 过渡数据:生存分析

17.1 引 论

持续期限的经济计量模型是关于由一种已知状态转入另一种状态时经历的时间长度的模型，比如失业期限、生命长短或没有健康保险时段。在生物统计学里，处于一种状态的持续期限，也是著名的生存时间[1]（**lifetime**），而过渡时间称为死亡（**death**）；在运算学里，人们要经常研究物体比如灯泡与机器的寿命，有效寿命结束也就是转入无效寿命，这称之为失效时间（**failure time**）。在经济计量学里，状态（**state**）是对单个个体处于时间某一刻的分类，过渡[2]（**transition**）则是从一种状态变到另一种状态，时期长度或持续期限是指在某已知状态下所经历的时间。一个典型的回归例子是，较高失业救济金对失业时期平均长度或对脱离失业概率的影响。

这个专题的文献数量巨大，令人感到茫然，其产生的原因众多。第一，几种有关的分布函数是关注的焦点，要么对过渡概率建模，要么对持续期限建模。第二，存在多种可能抽样方案，而统计推断既依赖于持续期限又依赖于抽样方案。例如，有关失业人员持续期限数据的抽样方法，包括流量抽样（**flow sampling**）——对在某已知月份成为失业者的人员进行抽样；存量抽样（**stock sampling**）——对在某已知月份失业者进行抽样；不论就业状况如何，而对所有人员总体进行抽样。第三，处于持续期限的数据经常是被截尾的。这是对过渡而不是通常回归分析目标即平均期限进行建模的主要原因，其目的是因为一致估计过渡模型要求较弱分布假设。第四，过渡数据具有相当丰富的各种状态，诸如失业、部分就业、全日就业以及非劳动力，同时对已知个体而言，可能运用这些状态的多种过渡数据。第五，文献中出现各种不同风格、各具特色的不同统计应用领域。在生物统计学中，持续期限分析（**duration analysis**）或过渡分析（**transition analysis**）也称为生存分析（**survival analysis**）（生存时间的长度），在运筹学中称为失效时间分析（**failure time analysis**）（某个研究对象比如灯泡或机器部件的故障时间长度），在人口学与精算研究中，则将

〔1〕 又称为寿命。——译者注

〔2〕 transition 在不同环境下有多种不同翻译术语，比如经济学中经常译成“转型”；而在数学和统计学中，译为“转移”、“转换”、“过渡”；这里译为“过渡”。——译者注

其称为生命表分析(**life table analysis**)(脱离状态对应于死亡)。在保险与事故理论中,它被称为风险分析(**hazard analysis**)。在社会科学应用中,包括惯犯,婚姻长短、选举间隔期限。

在本章,我们对通过流量抽样获得的单时期持续期限(**singe-spell duration**)数据阐述一些结果。一个经典例子是,对生存时间即从生到死的过渡进行建模,而且众多结果都来自生存分析与生命表分析。这是统计学中过渡分析最广泛研究的例子,而本章所述的生存分析方法,利用许多统计软件或微观经济计量软件来完成。本章以回归例子开始,概述由生存数据引发的问题。

17.3 节至 17.5 节阐述在没有回归元条件下的结果,因为在此情况下甚至会出现一些新概念。17.3 节引入基本持续期限数据概念,诸如风险、累积风险以及生存函数。17.4 节定义各种删失形式,这是持续期限分析普遍出现的新的复杂问题,因为完整时段不是总能被观测到的。例如,临床试验通常会在最后受试者死亡之前就结束。17.5 节阐述风险、累积风险(纳尔逊—奥伦估计量)以及生存函数(卡普兰—迈耶估计量)的非参数估计量,它们在独立删失条件下是一致的。

本章余下内容,再次在独立删失条件下,对回归模型加以推广。17.6 节阐述完全参数模型即著名威布尔模型的估计。对删失研究类似于对完全参数 Tobit 模型的那种研究。17.7 节给出某些重要的持续期限模型。不过,一种可供选择的半参数方法是对风险函数即以生存记载为条件的死亡概率进行建模。考克斯(Cox,1972)在他的原创性论文里,提出在相对较弱分布假设下,一致估计独立删失的比例风险函数方法。17.8 节阐述生存数据的标准模型——考克斯模型。与大多数横截面模型不同,生存模型的回归元比如失业持续期限模型中的失业救济金,对于已知人员来说,在受试者是可观测的时期内可能会变化。17.9 节详述含有时变回归元的模型。17.10 节阐述离散风险模型。而 17.11 节则给出一个实证例子。

后面两章将考察过渡建模的更复杂内容,这在已有教科书里很少被研究。这些内容包括不可观测异质性、多重时期以及多重目标。

17.2 罢工期限例子

考察由凯南(Kennan, 1985)、贾吉娅(Jaggia, 1991c)以及其他一些人曾经用过的罢工期限数据集合。关注的变量是,美国制造业从罢工开始时以天数测算的罢工期限。样本有 566 个罢工期限的完整(未删失)观测值。罢工的平均期限(dur) 为 43.6 天,中位数是 28 天左右。不过,在罢工开始之后 90 天,仍有 88 次罢工在坚持中。

我们可以证明,从图形上看,罢工期限信息是一个经验生存函数(**survival function**)。图 17.1 中的纵坐标轴表示,罢工在开始数天后仍在继续的比例。该图忽略日历时间,这意味着各次罢工的不同开始日期在建立图形时不起作用。正如人们所料,函数从 1 开始且单调衰退于 0,这显示所有罢工终将结束。

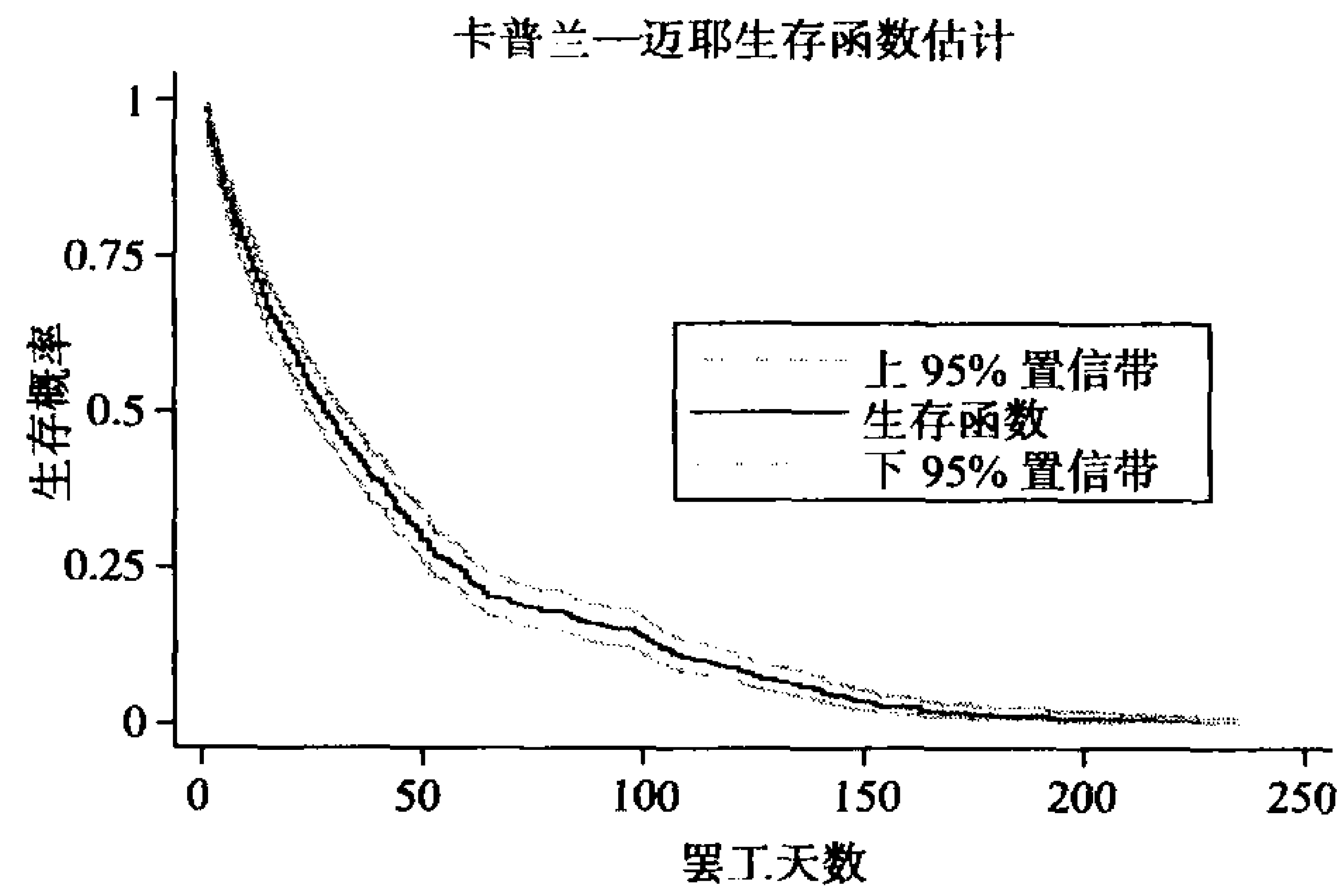


图 17.1 罢工期限：生存函数的卡普兰—迈耶估计。数据来自 1968～1976 年美国 566 次罢工的完整时期。

现在，引入回归元变量(z)，它测算偏离趋势水平——经济商业周期位置指示变量的程度。正的 z 值显示处于向上增长时期，而负的 z 值则正好相反。我们假定主要目标是检验平均罢工期限是处于周期前的[即$\partial(dur)/\partial z>0$]，还是处于周期后的[即$\partial(dur)/\partial z<0$]。一种继续研究的简单方法是，通过 $\ln(dur)$对 z 线性回归，来对 $\ln(dur)$的条件期望进行建模。如果人们想要对 dur 与 z 之间是否存在正相关或负相关加以检验，这样做就符合此目的。

可是，我们可能对罢工条件概率建模感兴趣。这个目标可利用具有结果为 0 或 1 的变量的二项回归来达到。不过，一旦控制 z，假定关注目标是在 t 天仍继续坚持的罢工在第 $t+1$ 天将要结束的概率，或是继续进行罢工将要结束的条件概率，并作为罢工时间长度的函数；那么，与生存分析相比，前面提及的回归方法则显得更缺乏方向性和效率，而生存分析还拥有另外一个优势，即生存分析可处理删失期限。下一节将考察用于生存分析的一些统计概念。

17.3 基本概念

某一种状态的持续期限是一个非负随机量，记为 T，就经济数据而言，T 经常是一个离散随机变量。为便于解释基本概念，我们关注连续情况，本章稍后给出离散情况。

17.3.1 *生存函数、风险函数以及累积风险函数*

T 的累积分布函数(**cumulative distribution function**)记为 $F(t)$，其密度函数是 $f(t)=dF(t)/dt$。那么，持续期限或时期长度小于 t 的概率是：

$$\begin{aligned} F(t) &= \Pr[T \leqslant t] \\ &= \int_0^t f(s)ds \end{aligned} \tag{17.1}$$

与 cdf 互补的一个有关概念是，持续期限等于或大于 t 的概率，它称为生存函

数(**survivor function**),定义如下:

$$S(t) = \Pr[T > t] = 1 - F(t) \tag{17.2}$$

式(17.1)中 cdf 的定义等同于遵从卡尔布弗莱舍与普伦蒂斯(Kalbfleisch and Prentice, 2002)的通常定义。在持续期限文献中,其他一些学者比如兰开斯特(Lancaster, 1990)反而定义 $F(t)=\Pr[T<t]$,从而 $S(t)=\Pr[T\geqslant t]$,因为如下定义的风险函数是以 $T\geqslant t$ 为条件而不是以 $T>t$ 为条件。在离散情况下,17.3.2 节考察过渡发生的准确时间,并且所用定义将取差分。

由于cdf是从0开始单调递增的,所以生存函数则是从1到0单调下降。所有个体都终将冒离开状态的危险,因而 $S(\infty)=0$。否则,$S(\infty)>0$,从而持续期限分布被称为不完美的。完整时期长度的样本均值是积分$\int_0^\infty S(u)du$。为了得到这一结果,使用:

$$\int_0^\infty uf(u)du = \int_0^\infty udF(u) = uF(u)\Big|_0^\infty - \int_0^\infty F(u)du$$

由于下 $F(\infty)=1$ 且 $F(0)=0$,由此可得:

$$\mathrm{E}[T] = \int_0^\infty (1-F(u))du = \int_0^\infty s(u)du \tag{17.3}$$

平均持续期限等于生存曲线以下的面积。

另一个重要概念是风险函数(**hazard function**),它是以一直生存到时间 t 为条件离开状态的瞬时概率。这被定义成:

$$\lambda(t) = \lim_{\Delta t\to 0}\frac{\Pr[t\leqslant T<t+\Delta t \mid T\geqslant t]}{\Delta t} = \frac{f(t)}{S(t)} \tag{17.4}$$

容易证明,该风险等于对数生存函数的变化率:

$$\lambda(t) = -\frac{d\ln(S(t))}{dt}$$

风险 $\lambda(t)$设定了 T 分布。特别地,对 $\lambda(t)$进行积分,并使用 $S(0)=1$,可以证明:

$$S(t) = \exp\left(-\int_0^t \lambda(u)du\right) \tag{17.5}$$

在过渡回归分析中,条件风险率 $\lambda(t|\mathbf{x})$是关注的核心内容。这一点可与更标准的回归方法形成一种对比,即标准回归方法里,条件均值函数 $\mathrm{E}[T|\mathbf{x}]$是其关注焦点。后面这一方法具有不利条件,在实际应用中,持续期限经常被删失。

最后一个有关的函数是累积风险函数(**cumulative hazard function**)或综合风险函数[1](**integrated hazard function**):

[1] 又称为积分风险函数。——译者注

$$\Lambda(t) = \int_0^t \lambda(s)ds \tag{17.6}$$
$$= -\ln S(t)$$

其中，最后一个等式运用了式(17.5)。当 $S(\infty)=0$ 时，有 $\Lambda(\infty)=\infty$。与风险函数相比，由于累积风险更能准确地得到估计，所以它是关注的内容。

对 T 分布的任何选取而言，可以证明，变换 $\Lambda(T)$ 都是单位指数分布，而 $\ln\Lambda(T)$ 服从极值分布，这为对模型设定进行检验提供了基础，参见 18.7.2 节。

关于非负连续随机变量 T 的各种有关函数，已由表 17.1 概括归纳。

表 17.1　生存分析：重要概念的定义

函数	符号	定义	关系
密度	$f(t)$		$f(t)=\frac{dF(t)}{dt}$
分布	$F(t)$	$\Pr[T\leqslant t]$	$F(t)=\int_0^t f(s)ds$
生存函数	$S(t)$	$\Pr[T>t]$	$S(t)=1-F(t)$
风险	$\lambda(t)$	$\lim_{h\to 0}\frac{\Pr[t\leqslant T<t+h\mid T\geqslant t]}{h}$	$\lambda(t)=\frac{f(t)}{S(t)}$
累积风险	$\Lambda(t)$	$\int_0^{\infty}\lambda(s)ds$	$\Lambda(t)=-\ln S(t)$

有时，还会运用其他一些函数，像最著名的拉普拉斯变换 $L(s)=\mathrm{E}[\exp(-sT)]$，$s>0$，它是那种将随机变量 T 限制为正的矩生成函数的一种变形。

17.3.2　*离散数据*

一种非常普遍的情形是，持续期限以区间形式加以度量。例如，数据可能显示过渡发生在某个特定周里，但并不知道该周的某个准确时间。在此类情况下，过渡时间被称为分组的，并假定区间之内的风险为常值。离散时间风险模型将研究这种数据。

讨论起点是，将离散时间风险函数定义成，已知一直生存到时间 t_j，在离散时间 t_j 过渡的概率是$(j=1,2,\cdots)$：

$$\lambda_j = \Pr[T=t_j \mid T\geqslant t_j] \tag{17.7}$$
$$= f^d(t_j)/S^d(t_{j_})$$

其中，上标 d 表示离散的，而 $S^d(a_-)=\lim_{t\to a} S^d(t_j)$，由于在形式上 $S^d(t)$ 等于 $\Pr[T>t]$ 而不是 $\Pr[T\geqslant t]$，所以要做调整，并且上标"d"表示离散。

从风险函数中，可递归地获得离散时间生存函数(**discrete-time survivor function**)，即：

$$S^d(t) = \Pr[T\geqslant t] \tag{17.8}$$
$$= \prod_{j\mid t_j\leqslant t}(1-\lambda_j)$$

例如，$\Pr[T>t_2]$ 等于在时间 t_1 没有过渡的概率乘以刚好在 t_2 之前以生存为条件的、在时间 t_2 没有过渡的概率，所以 $\Pr[T>t_2]=(1-\lambda_1)\times(1-\lambda_2)$。函数 $S^d(t)$ 在

t_j 处是一个递减阶梯函数,其阶梯步长在 t_j 出现,$j=1,2,\cdots$。

离散时间累积风险函数(**discrete-time cumulative hazard function**)是:

$$\Lambda^{d}(t) = \sum_{j|t_j \leqslant t} \lambda_j \tag{17.9}$$

利用式(17.7),得出时期在 t_j 结束时的离散概率是 $\lambda_j S^{d}(t_j)$。

能够将连续情况与离散情况联合起来。于是,运用乘积积分定义生存函数,在离散情况下,乘积积分简化成普通乘积式(17.8),而在连续情况下,乘积积分简化成普通积分的指数形式。参见卡尔布弗莱舍和普伦蒂斯(Kalbfleisch and Prentice, 2002,第 10 页)或兰开斯特(Lancaster, 1990,第 10~12 页)。

由于过程生成过渡本质上是离散的,所以就产生了离散持续期限数据。可是,大多数情况下,基本过程是连续的,只是数据以离散方式观测到。例如,人们可能知道时期在某个星期或某个月份结束了,却不知道是哪天或几点结束的。这种数据有时统称为分组数据(**grouped data**)。离散数据能系统地表述如下。设时间被 $k+1$ 个区间$[a_0,a_1)$,$[a_1,a_2)$,$\cdots$,$[a_{k-1},a_k)$, $[a_k,a_\infty)$分割。离散时间持续期限 $T=t_j$ 代表在区间$[a_{j-1},a_j)$内发生过渡,也就是说,在时间 a_{j-1} 或稍后出现过渡。一种习惯做法是,将离散数据处理成由分组而导致的,因此对过渡可用连续时间形式加以建模,然后通过分组做出必要调整。进一步讨论将由 17.10 节给出。

17.4 删　失

通常,生存数据是被删失的,因为某些时期不能完全被观测到。也就是说,仅仅知道寿命位于某些区间之中。举一个例子,除观测到失业的完整时期长度之外,数据可能来自对当前失业者的调查,因此,只有失业不完整时期的长度才能被观测到。

17.4.1 *删失机制*

在实际应用中,数据可能是右删失的、左删失的或者区间删失的。对于右删失(**right censoring**)或从上面删失来说,我们观测到从时间 0 到删失时间 c 的时期。一些时期将到这个时间(完整时期)结束,而另一些时期则是不完整的,但我们知道的内容是,它们在区间(c,∞)内的某个时刻结束。当知道某些时期在区间$(0,c)$内的某个时间结束,但不知道其准确时间时,就出现左删失(**left censoring**)或从下面删失。经典 Tobit 模型就是一个例子,其中某些时期上的数据丢失,而且删失时间是未知的。当观测到完整时期长度,但仅仅是以区间形式比如$[t_1^*,t_2^*)$出现,便发生区间删失。

生存分析文献关注于右删失。即使有这种限制,但仍有众多原因可能导致删失,删失包括随机删失、第 I 类删失以及第 II 类删失。

随机删失或外生删失意指,样本中的每个个体具有完整持续期限 T_i^* 与删失时间 C_i^*,它们之间相互独立。若时期在删失时间之前结束,我们观测到完整持续期限 T_i^*;若时期在删失时间之后结束,我们观测到删失时间 C_i^*。此外,还可以知

道删失发生与否。观测数据(t_1,δ_1)，(t_2,δ_2)，…，(t_N,δ_N)是随机变量的实现值：

$$T_i=\min(T_i^*,C_i^*) \tag{17.10}$$
$$\delta_i=\mathbf{1}[T_i^*<C_i^*]$$

当A发生时，指示函数$\mathbf{1}[A]=1$，否则$\mathbf{1}[A]=0$。注意，当完整时期是可观测的，$\delta_i=1$，否则为0。随机删失可能由下述原因引起，诸如由下列情况引致的随机失效：个体随机地从研究中退出或研究终止。

第I类删失(**type I censoring**)意指，当期限在某个固定已知删失时间比如t_{c_i}之上时被删失。例如，灯泡样本对所有对象都具有共同开始时间，并且对不超过5 000小时的加以检验。因而，在研究终止时，某些对象的失效时间或持续期限是已知的，但其他目标仍没有"失效"。可以认为，它们的寿命是右删失的。这是随机删失满足$C_i^*=t_{c_i}$的一种特殊情况。经典Tobit模型是关于连续$(-\infty,\infty)$区间上随机变量从下面删失的第I类删失的一个例子。

17.4.2 独立（非信息）删失

在存在删失条件下，为了使标准生存分析方法成为有效的，就要求删失机制是那种具有独立（非信息）删失的。这意味着，C^*分布的参数不涉及持续期限T^*分布参数的信息。于是，人们将删失指示变量δ处理成外生的，而且若关注内容在于持续期限参数，则不必对删失机制进行建模。

对于删失数据(t,δ)来说，未删失观测值以概率：

$$\Pr[T=t,\ \delta=1]=\Pr[T=t|\delta=1]\times\Pr[\delta=1]$$

被观测到。若删失机制是独立的，则$\Pr[T=t|\delta=1]=\Pr[T=t]$。如果删失机制是非信息的，那么$\Pr[\delta=1]$这项可从似然函数中去掉，因为它不涉及$T$分布的参数。类似地，对于删失观测值来说：

$$\Pr[T=t,\ \delta=0]=\Pr[T\geqslant t|\delta=0]\times\Pr[\delta=0]$$

在独立删失条件下，$\Pr[T\geqslant t|\delta=0]=\Pr[T\geqslant t]$，而在非信息删失条件下，可省略$\Pr[\delta=0]$。一旦与上述讨论结合起来，关注密度简化成，当$\delta=1$时$\Pr[T=t]$，当$\delta=0$时$\Pr[T\geqslant t]$。

当引入回归元$\mathbf{x}$时，对于T^*与C^*来说，可能出现随同一回归元而变化。这时，C^*参数再次不涉及T^*参数的信息。更简单地讲，在任何给定时点上，不一定发生删失，因为给定$\mathbf{x}$时，实验者具有非常高或非常低的失效风险。

第II类删失(**type II censoring**)意指，N个实验者的观测值，在第p个失效之后将终止。从而，仅有p个最短时期的持续期限才是完全可观测到的，而余下来的$N-p$个则在$C_i^*=t_{(p)}$处被删失，即第p个最短时期的完整时期。例如，临床试验可能在p个病人去世之后停止。

随机删失、第I类删失以及第II类删失都是独立删失的一些例子。更正式的研究，由卡尔布弗莱舍和普伦蒂斯(Kalbfleisch and Prentice, 2002，第194～196页)给出。

17.5 非参数模型

本节研究生存函数的非参数估计,就描述目的而言,这些方法极为有用。在考虑引入回归元之前,在直观上人们时常要了解原来(无条件)风险或生存函数的形状。运用罢工持续期限例子就可阐明这一点。

我们阐述,存在独立删失条件下的生存函数、风险函数以及累积风险函数的估计量。密度本身的非参数因为很难通过删失而引入,故而不便考虑;更为重要的是,与密度相比,生存函数及风险函数更容易解释。

回归元没有被包括进来。若关注内容仅仅是回归元的几个重要值,诸如各种不同处理制度或处理水平,则人们可在每个重要值上获得各自的非参数估计,并对它们加以比较。在经济学应用中,很少出现这种情况,而是需要拥有回归元结构更丰富的模型,这将在 17.6 节至 17.10 节讨论。

我们关注离散持续期限,比如生命表数据,故要用到 17.3.3 节的离散时间公式。例如,考虑特定年龄与性别的 N_0 个一组人。对他们跟踪数年。在第 1 年末,此组存在 N_1 个人,而 N_1-N_0 个人在最初组中要么因为死亡要么由于其他缘故而丢失(删失)。在随后一年,此组人数为 N_2-N_1,等等。这种生命表数据能被用于构造在没有任何先验参数假设条件下的离散时间生存函数。

17.5.1 *非参数估计*

就没有删失情况而言,生存函数的一个明显估计量是 1 减去样本累积分布函数。于是,$\hat{S}(t)$等于持续期限大于 t 的样本中的时期个数被样本量 N 除。这是在离散失效时间处具有跳跃性的阶梯函数;参见图 17.1。给定式(17.13),这个估计量的一种可供选择等价表述形式,在存在独立删失条件下保持一致性。

设 $t_1<t_2<\cdots<t_j<\cdots<t_k$ 表示样本量为 N 的样本中时期可观测的离散失效时间,$N\geqslant k$。定义 d_j 为在时间 t_j 结束的时期数。由于数据是离散的,所以 d_j 可能大于 1。一些时期可能是不完整观测的。定义 m_j 为区间$[t_j,t_{j+1})$中右删失时期数。若假定删失机制是独立删失的,则对于处于$[t_j,t_{j+1})$之中的删失时期,只知道其失效时间大于 t_j。如果一些时期尚未失效或没有删失,它们就处于失效风险之中。把 r_j 定义成等于在时间 t_{j-},即刚好在时间 t_j 之前处于风险之中的时期数。于是,$r_j=(d_j+m_j)+\cdots+(d_k+m_k)=\sum_{l|l\geqslant j}(t_l+m_l)$。注意到,$r_1=N$。总之:

$$
\begin{aligned}
d_j &= \text{在时间 } t_j \text{ 结束的时期数} \qquad (17.11)\\
m_j &= \text{在}[t_j,t_{j+1})\text{ 内的删失时期数}\\
r_j &= \text{在时间 } t_{j-} = \sum_{l|l\geqslant j}(d_l+m_l) \text{ 的风险时期数}
\end{aligned}
$$

可运用 17.3.2 节的离散时间公式。由于 $\lambda_j=\Pr[T=t_j\mid T\geqslant t_j]$,所以风险函数的一个明显估计量是,在时间 t_j 结束的时期数被在时间 t_{j-} 处于失效风险时期数除,或:

$$\hat{\lambda}_j = \frac{d_j}{r_j} \tag{17.12}$$

离散时间生存函数已由式(17.8)定义。生存函数的卡普兰—迈耶估计量或乘积极限估计量是样本的类似形式：

$$\hat{S}(t) = \prod_{j|t_j \leqslant t} (1 - \hat{\lambda}_j) = \prod_{j|t_j \leqslant t} \frac{r_j - d_j}{r_j} \tag{17.13}$$

这是一个递减阶梯函数，在每个离散失效时间处都有一个跳跃。可以证明，卡普兰—迈耶估计量是非参数的 MLE[参见卡尔弗莱舍和普伦蒂斯(Kalbfleisch and Prentice, 2002,第 14～16 页)]。

在没有删失的情况下，式(17.13)的 $\hat{S}(t)$ 可简化成 $\hat{S}(t) = r/N$，在时间 t 处于风险的时期数仍被样本量除，即 1 减去经验 cdf。为了理解这一点，注意到 $r_j - d_j = r_{j+1}$，当 $m_j = 0$ 时，在时间 j 的处于风险时期数小于在时间 j 时的终止时期数，这等于在时间 $j+1$ 的处于风险时期数。于是，式(17.13)变成 $\hat{S}(t) = \prod_{j|t_j \leqslant t} r_{j+1}/r_j$，从而简化成 r/r_1，其中，$r_1 = N$。

离散时间累积风险函数已由式(17.9)定义。累积风险函数的纳尔逊—奥伦(**Nelson - Aalen**)估计量是一个明显的样本类似形式：

$$\hat{\Lambda}(t) = \prod_{j|t_j \leqslant t} \hat{\lambda}_j = \prod_{j|t_j \leqslant t} \frac{d_j}{r_j} \tag{17.14}$$

这个估计量也可通过 $\tilde{S}(t_j) = \exp(-\hat{\Lambda}(t))$ 用于估计生存函数，而在连续情况下，则利用等式 $S(t) = \exp(-\Lambda(t))$ 来进行。

举一个例子，假定最初有 80 年观测值，在时间 t_1 有 6 个失效，在 $[t_1, t_2)$ 内有 4 个时期删失，在时间 t_2 有 5 个失效，在 $[t_2, t_3)$ 内有 3 个时期删失，在时间 t_3 有 2 个失效，在 $[t_3, t_4)$ 内有 1 个时期删失等。于是，当 $t \leqslant t_3$ 时，生存函数与累积风险估计由表 17.2 给出。

表 17.2　风险率与生存函数的比较：例子[a]

j	r_j	d_j	m_j	$\hat{\lambda}_j = d_j/r_j$	$\hat{\Lambda}(t_j)$	$\hat{S}(t_j)$
1	80	6	4	6/80	6/80	(1−6/80)
2	70	5	3	5/70	6/80+5/70	(1−6/80)×(1−5/70)
3	62	2	1	2/62	$\hat{\Lambda}(t_2)$+2/62	$\hat{S}(t_2)$×(1−2/62)
4	—	—	—	—		

[a] 在时间 t_j，r_j 表示处于风险之中的观测值数目，d_j 表示死亡(失效)数目，m_j 表示缺失时期(被删失的)数目，$\hat{\lambda}_j$ 表示估计风险率，$\hat{\Lambda}(t_j)$ 表示估计累计风险，$\hat{S}(t_j)$ 表示估计生存函数。

结数据(**tied data**)意指，多重失效在一个特定时点上发生。一种普遍的假定是，由于分组而出现结数据，并不是因为过程生成了真实离散结。风险估计值 $\hat{\lambda}_j = d_j/r_j$ 假定所有终止都同时发生在时间 t_j。实际上，终止可在区间 $[t_j, t_{j+1})$ 内以累进方式发生，删失也可能在此区间上以累进方式出现。那么，r_j 对区间 $[t_j, t_{j+1})$ 处

于风险中的实验者数量平均来说会高估。在生命表分析中,标准修正法是用 $d_j/(r_j-m_j/2)$ 代替 $\hat{\lambda}_j=d_j/r_j$,对于 $\hat{S}(t)$、$\hat{\Lambda}(t)$ 等公式,可做出类似变化。还可提出其他一些修正法。

绝大多数生存分析方案都会产生基本的卡普兰—迈耶图形和表。表 17.3 给出罢工数据这种输出的摘要,并补充前面由图 17.1 给出的图标。

表 17.3 罢工持续期限:卡普兰—迈耶生存函数估计

天数	开始总数	失效	生存函数	标准误差
1	566	10	0.982 3	0.005 5
2	556	21	0.945 2	0.009 6
3	535	16	0.917 0	0.011 6
4	519	17	0.886 9	0.013 3
5	502	18	0.855 1	0.014 8
6	484	9	0.839 2	0.015 4
7	475	12	0.818 0	0.016 2
8	463	12	0.796 8	0.016 9
⋮	⋮	⋮	⋮	⋮
13	411	11	0.706 7	0.019 1
14	400	11	0.687 3	0.019 5

17.5.2 *非参数估计的置信带*

风险函数的 $\hat{\lambda}_j=d_j/r_j$ 估计值是相当不连续的,尤其是对于大 t 而言,因为那样 r_j 相对于 d_j/r_j 来说变得很少。在画出风险估计与时间的图形之前,首先利用非参数回归方法对风险估计进行光滑处理,这样做在形式上很有用,参见 9.5 节。

生存函数与累积风险函数都更加光滑,一种标准做法是,将这些函数与时间变化画出图形,并且反映抽样变异的置信带。有几种方法估计这些置信带。我们给出的公式都是运用 STATA 的。

对于生存函数的卡普兰—迈耶估计来说,一种普通做法是,运用方差的格林伍德估计值:

$$\hat{V}[\hat{S}(t)]=\hat{S}(t)^2\sum_{j|t_j\leqslant t}\frac{d_j}{r_j(r_j-d_j)}$$

被报告出来的 $S(t)$ 置信区间经常建立在 $\ln(-\ln\hat{S}(t))$ 而不是 $\hat{S}(t)$ 的基础上,因为这种变换确保了置信区间位于生存函数范围内,即位于 0~1 之间。由此变换,得出 $100(1-\alpha)\%$ 置信区间为:

$$S^{d}(t)\in(\hat{S}(t)\exp^{(-z_{\alpha/2}\hat{\sigma}(t))},\ \hat{S}(t)\exp^{(z_{\alpha/2}\hat{\sigma}(t))})\qquad(17.15)$$

其中,$\sigma(t)$ 表示 $\ln(-\ln\hat{S}(t))$ 的标准差,它是利用:

$$\hat{\sigma}_s^2(t)=\frac{\sum_{j|t_j\leqslant t}d_j/(r_j(r_j-d_j))}{[\sum_{j|t_j\leqslant t}\ln((r_j-d_j)/d_j)]^2}$$

估计出来的。

对于累积风险函数的纳尔逊—奥伦估计量，方差估计值是：

$$\hat{V}[\hat{\Lambda}(t)] = \sum_{j|t_j\leq t} \frac{d_j}{r_j^2}$$

由变换 $\ln \hat{\Lambda}(t)$，得出累积风险的 $100(1-\alpha)\%$置信区间：

$$\Lambda(t) \in [\hat{\Lambda}(t)\exp(-z_{\alpha/2}\hat{\sigma}_{\Lambda}(t)),\ \hat{\Lambda}(t)\exp(z_{\alpha/2}\hat{\sigma}_{\Lambda}(t))] \tag{17.16}$$

其中，$\hat{\sigma}_{\Lambda}(t)$表示 $\ln \hat{\Lambda}(t)$的标准差，也可利用

$$\hat{\sigma}_{\Lambda}^2(t) = \hat{V}[\hat{\Lambda}(t)/\hat{\Lambda}^2(t)]$$

估计出来。

17.6 参数回归模型

我们通过概述起基准作用的两个分布的性质开始讨论。对于持续期限数据来说，考察一些标准回归模型。

17.6.1 指数分布和威布尔分布

正常参数起点是指数，因为纯泊松点过程具有服从指数分布的持续期限，参见兰开斯特(Lancaster, 1990，第 86 页)。指数持续期限分布具有常值风险率 γ，它不随 t 而变化，拥有指数的无记忆性质。由式(17.5) 可得，$S(t) = \exp(-\int_0^t \gamma du) = \exp(-\gamma t)$。密度是 $f(t) = -S'(t) = \gamma\exp(-\gamma t)$，而累积风险 $\Lambda(t) = -\ln S(t) = \gamma t$ 关于 t 是线性的。

在实际应用中，因指数分布是单参数分布，故表现出极强的约束性。经济计量学中普遍使用的推广形式是威布尔分布(**Weibull distribution**)。表 17.4 列出威布尔分布与指数分布的密度、其他分布函数以及各阶矩，这里特殊情况为 $\alpha=1$。由 17.5 表给出的函数 $\Gamma(\cdot)$是一个伽玛函数。

表 17.4 指数分布与威布尔分布：pdf、cdf、生存函数、风险、累积风险、均值以及方差

函数	指数	威布尔
$f(t)$	$\gamma\exp(-\gamma t)$	$\gamma\alpha t^{\alpha-1}\exp(-\gamma t^{\alpha})$
$F(t)$	$1-\exp(-\gamma t)$	$1-\exp(-\gamma t^{\alpha})$
$S(t)$	$\exp(-\gamma t)$	$\exp(-\gamma t^{\alpha})$
$\lambda(t)$	γ	$\gamma\alpha t^{\alpha-1}$
$\Lambda(t)$	γt	γt^{α}
$E[T]$	γ^{-1}	$\gamma^{-1/\alpha}\Gamma(\alpha^{-1}+1)$
$V[T]$	γ^{-2}	$\gamma^{-2/\alpha}[\Gamma(2\alpha^{-1}+1)-[\Gamma(\alpha^{-1}+1)]^2]$
γ,α	$\gamma>0$	$\gamma>0,\ \alpha>0$

表 17.5 标准参数模型及其风险函数与生存函数[a]

参数模型	风险函数	生存函数	类型
指数	γ	$\exp(-\gamma t)$	PH, AFT
威布尔	$\gamma\alpha t^{\alpha-1}$	$\exp(-\gamma t^{\alpha})$	PH, AFT
广义威布尔	$\gamma\alpha t^{\alpha-1}S(t)^{-\mu}$	$[1-\mu\gamma t^{\alpha}]^{1/\mu}$	PH
冈珀茨	$\gamma\exp(-\alpha t)$	$\exp(-(\gamma/\alpha)(e^{\alpha t}-1))$	PH
对数正态	$\dfrac{\exp(-(\ln t-\mu)^2/2\sigma^2)}{t\sigma\sqrt{2\pi}[1-\Phi((\ln t-\mu)/\sigma)]}$	$1-\Phi((\ln t-\mu)/\sigma)$	AFT
对数逻辑斯蒂	$\alpha\gamma^{\alpha}t^{\alpha-1}/[(1+(\gamma t)^{\alpha})]$	$1/[(\gamma t)^{\alpha}]$	AFT
伽玛	$\dfrac{\gamma(\gamma t)^{\alpha-1}\exp[-(\gamma t)]}{\Gamma(\alpha)[1-I(\alpha,\gamma t)]}$	$1-I(\alpha,\gamma t)$	AFT

[a] 对于冈珀茨(Gomportz)模型,除 $\infty<\alpha<\infty$ 之外,所有参数都被限制成正的。

威布尔模型具有风险函数 $\lambda(t)=\gamma\alpha t^{\alpha-1}$,当 $\alpha>1$ 时,它是单调递增的;而当 $\alpha<1$ 时,它是单调递减的。这是比例风险(PH)族的特殊情况,参见 17.7.1 节,在此情况下,$\lambda(t)$ 因子归入仅仅依赖于 t、$\lambda_0(t)$ 的基准成分之中,第二项(即 γ)能被参数化成唯一协变量的函数。图 17.2 显示了 $\gamma=0.01$ 与 $\alpha=1.5$ 时威布尔分布的性质。如同具有持续期限数据的情况,密度是向右偏斜的。生存曲线的形状是许多各种不同分布体现出来的共同形式,要直接分辨各种估计生存曲线则很困难。在威布尔例子中,风险函数是递增的,这是因为 $\alpha>1$。其他参数模型能具有各种不同形状的风险函数,包括单调递增的形状、单调递减的形状、U 形状以及反 U 形状。

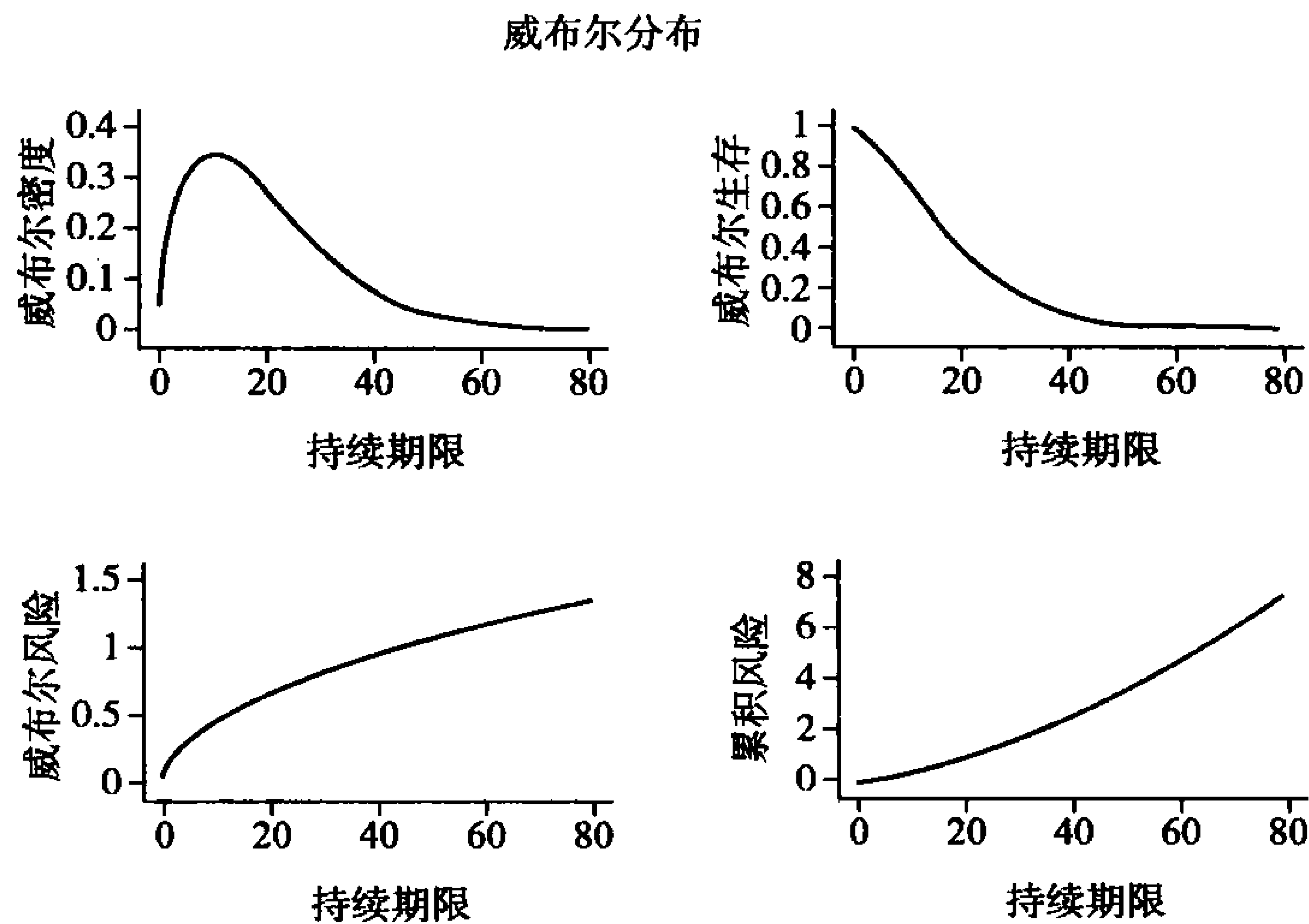


图 17.2 威布尔分布:对应于时间 $\gamma=0.01$ 与 $\alpha=1/5$ 时的密度函数、生存函数、风险函数、累积风险函数的散点图。

在实际应用中,很难被准确地估计出风险函数,尤其对右边尾部。而累积风险 $\Lambda(t)$ 则可更准确被估计出,并对各类模型进行辨别成为可能。一种更好的方法是,

画出 $\ln \Lambda(t)$ 与 $\ln t$ 的图形，其原因在于威布尔模型 $\ln \Lambda(t)=\ln \gamma+\alpha \ln t$ 关于 $\ln t$ 是线性的，其斜率为 α。

17.6.2 *某些参数模型*

深受欢迎的参数模型包括指数模型、威布尔模型、冈珀茨模型、对数正态模型、对数逻辑斯蒂模型以及伽玛模型。这些模型的风险与生存函数已由表 17.5 给出。

对于伽玛模型，$\Gamma(\alpha)=\int_0^{\infty} e^{-t} t^{\alpha-1} dt$ 是伽玛函数，而 $I(\alpha, \gamma t)$ 是不完整伽玛函数，其中，$I(a, x)=\int_0^{x} e^{-t} t^{\alpha-1} dt / \Gamma(a)$，$0<I(a, x)<1$。

广义威布尔模型是由穆达尔卡、斯里瓦斯塔瓦和科勒(Mudholkar, Srivastava and Kollia, 1996)提出的。威布尔模型通过引入其他形状参数，克服了对那个模型的重要限制，并促使风险函数拥有更加灵活的形状。当 $\mu \rightarrow 0$ 时，通过取极限获得威布尔模型。由表 17.5 知道：

$$\ln \lambda(t)=\ln(\gamma\alpha)+(\alpha-1)\ln t-\mu \ln S(t)$$

由于 $\partial \ln S(t)/\partial t<0$，所以当 $\mu>0$ 且 $\alpha>1$ 时，该方程右边关于 t 是递增的。当 $\alpha \leqslant 1$ 且 $\mu<0$ 时，风险函数是单调递减的。当 $\alpha>1$ 且 $\mu<0$ 时，风险函数具有两种成分，其中一种是关于 t 递减的，而另一种关于 t 则是递增的。因此，两个合并能生成单峰或 U 形风险函数。所以，广义威布尔模型具有潜在灵活性，是一种有用的函数形式。

冈珀茨模型类似于威布尔模型，因为它在(当 $\alpha>0$ 时)单调递增或者($\alpha<0$ 时)单调递减(当 $\alpha=0$ 时)特殊情况下作为指数模型。冈珀茨模型是死亡数据方面的一个优秀模型，它在生物统计学中的应用，比在经济计量学中的应用更加广泛。

对数正态分布具有倒置形的浴缸风险函数，即它首先随 t 变化而增大，然后随 t 变化而递减。当 $\alpha>1$ 时，对数逻辑斯蒂也是如此。对于拥有该性质的持续期限数据来说，很明显，与指数模型威布尔以及冈珀茨模型相比，这些模型更为合适。

另一些参数模型，包括基于雷利和梅卡姆(Rayleigh and Makeham)分布的模型、逆高斯逐段连续风险函数、广义伽玛模型(Lawless, 1982)均可嵌入到一种作为特殊情况的伽玛与威布尔模型中。卡尔布弗莱舍和普伦蒂斯(Kalbfleisch and Prentice, 2002，第 3 章)与兰开斯特(Lancaster, 1990，第 3 章)都详细地阐述了许多参数模型。

一般地讲，一些分布都是两个参数的分布。回归元是通过令 $\gamma=\exp(\mathbf{x}'\boldsymbol{\beta})$ 引入的，并设 α 为常值，但对于对数正态模型，则是令 $\mu=\mathbf{x}'\boldsymbol{\beta}$，并设 σ^2 为常值。

为了获得一致参数估计，并利用广泛的参数模型，参数建模的主要问题是对模型正确设定的依赖性。绝大多数模型被分类归入 PH 模型(表 17.5 中的前四个)或加速失效时间模型(表 17.5 中的前两个与后三个模型)之中。归属于这两类模型的是威布尔模型，它在经济学中的应用相当广泛。特别是，当经济应用中有许多观测值可利用时，另一种广泛使用的模型是分段常数风险模型，它是 PH 模型的一种特殊情况。

17.6.3 极大似然估计

我们现在利用 ML 与最小二乘方法,考察具有独立删失或非信息删失的完全参数分析。由于参数模型建立在连续分布基础之上,所以可使用连续持续期限公式。假定回归元是时不变的,而时变回归元的情况参见 17.9 节。

设 T^* 表示没有删失的持续期限,条件密度为 $f(t|\mathbf{x},\boldsymbol{\theta})$,其中,$\boldsymbol{\theta}$ 表示 $q\times1$ 维参数向量,$\mathbf{x}$ 是回归元,它随不同实验者而变化,但对一个给定实验者来说,却并不随时期变动而变化。由于存在删失,故估计颇为复杂。于是,观测到的持续期限 t 可能是不完整时期长度,数据被揭示删失存在的变量所扩大,这里,删失被假定成非信息的。

由 17.4.2 节,研究类似于对 Tobit 模型所做出的那样。对于未删失观测值,对似然函数的贡献是 $f(t|\mathbf{x},\boldsymbol{\theta})$。就右删失观测值而言,只知道大于 t 的持续期限,因此它对似然函数的贡献是:

$$\begin{aligned}\Pr[T>t]&=\int_t^{\infty}f(u|\mathbf{x},\boldsymbol{\theta})du\\&=1-F(t|\mathbf{x},\boldsymbol{\theta})=S(t|\mathbf{x},\boldsymbol{\theta})\end{aligned}$$

其中,$S(\cdot)$表示生存函数,第 i 个观测值的密度被写成:

$$f(t_i|\mathbf{x}_i,\boldsymbol{\theta})^{\delta_i}S(t_i|\mathbf{x}_i,\boldsymbol{\theta})^{1-\delta_i}$$

其中,δ_i 表示右删失指示变量,即:

$$\delta_i=\begin{cases}1, & (\text{未删失})\\0, & (\text{右删失})\end{cases}$$

一旦取对数并求和,得到对对数似然

$$\ln \mathrm{L}(\boldsymbol{\theta})=\sum_{i=1}^{N}[\delta_i\ln f(t_i|\mathbf{x}_i,\boldsymbol{\theta})+(1-\delta_i)\ln S(t_i|\mathbf{x}_i,\boldsymbol{\theta})] \tag{17.17}$$

求极大值的 MLE $\hat{\boldsymbol{\theta}}$,其中,假设对不同 i 具有独立性。和式中第一项对应完整时期,而第二项则对应右删失时期。由于 $\ln S(t)=\Lambda(t)$,且 $\ln f(t)=\ln(\lambda(t)S(t))=\ln\lambda(t)+\ln S(t)$,所以此对数的值可用另一种方式利用风险与综合风险函数写成:

$$\ln \mathrm{L}(\boldsymbol{\theta})=\sum_{i=1}^{N}[\delta_i\ln\lambda(t_i|\mathbf{x}_i,\boldsymbol{\theta})+\Lambda(t_i|\mathbf{x}_i,\boldsymbol{\theta})] \tag{17.18}$$

倘若参数模型是通过设定风险率而不是 pdf 定义,就可运用这一结果。

这里可应用通常的估计理论。如果密度被正确设定,那么 MLE 服从 $\hat{\boldsymbol{\theta}}\overset{a}{\sim}\mathcal{N}[\boldsymbol{\theta},(-\mathrm{E}[\partial^2\ln L/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'])^{-1}]$,参见 5.7.3 节。不过,若密度被错误设定,则 MLE 是非一致的。一个值得注意的例外是,存在删失条件下的指数持续期限模型,为了一致性,仅仅需要正确设定条件均值函数,参见 5.7.3 节。然而,甚至对指数模型来说,若引入删失,则在错误设定下出现非一致性,而对其他参数持续期限模型来说,甚至在没有删失时,也出现非一致性。正如 Tobit 模型情况一样,参数

方法的主要弱点是缺乏稳健性。

对 ML 方法加以改进，以便允许估计具有删失其他类型的模型。就左删失而言，可以知道，时期长度至多为 t，从而其对似然贡献是 $\Pr[T^* < t] = \int_0^t f(s \mid \mathbf{x}, \boldsymbol{\theta}) ds = F(t \mid \mathbf{x}, \boldsymbol{\theta})$。

对于区间删失而言，可以知道，数据位于$[t_a, t_b)$之中，从而其对似然贡献是：$\Pr[t_a \leqslant T^* < t_b] = \int_{t_a}^{t_b} f(s \mid \mathbf{x}, \boldsymbol{\theta}) ds = S(t_b \mid \mathbf{x}, \boldsymbol{\theta}) - S(t_a \mid \mathbf{x}, \boldsymbol{\theta})$。

在经济应用中，使用的持续期限经常是区间删失的。例如，失业持续期限可能被分成数周与数月，而参数模型是连续分布的，比如威布尔模型。通常假定区间删失的效应充分小，以至于可以忽略区间删失。比如，某个人 2 个月后为失业者，但 3 个月之后不再是失业者了，可以将此处理成拥有确切 3 个月失业时期，而不能处理成 2～3 个月范围的时期。

17.6.4 *似然函数成分*

倘若数据是持续期限形式的一种混合形式，即以前面曾经提及方式出现的完整数据、截尾或删失数据的混合体，则参数形式所设定模型的极大似然估计就要求人们建立其似然函数。[兰开斯特(Lancaster, 1979)已经提出，适合于失业持续期限背景下三种不同数据形式的各种似然表达式。]每一种类型观测值都成为似然函数的一项，而整个似然函数则是对如下各项以适当乘积形成的[参见科林和莫斯伯格(Klein and Moeschberger, 1997，第 66 页)]：

完整持续期限：	$f(t)$
在 t_L 处左截尾($t \geqslant t_L$)：	$f(t)/S(t_L)$
在 t_{C_L} 处左删失：	$1 - S(t_{C_L})$
在 t_{C_R} 处右删失：	$S(t_{C_R})$
在 t_{C_R} 处右截尾($t \leqslant t_R$)：	$f(t_R)/[1 - S(t_R)]$
在 t_{C_L}、t_{C_R} 处区间删失：	$S(t_{C_L}) - S(t_{C_R})$

17.6.5 *威布尔 MLE 例子*

威布尔分布已由 17.6.1 节给出了详细阐述。其风险函数是 $\lambda(t) = \gamma \alpha t^{\alpha-1}$，其中，$\alpha > 0$ 且 $\gamma > 0$。

回归元可通过多种方式引入，但通常设定 $\gamma = \exp(\mathbf{x}'\boldsymbol{\beta})$，这样做确保 $\gamma > 0$，而 α 并不随回归元而变化。[不过，一些方法设定 $\gamma = \exp(-\mathbf{x}'\boldsymbol{\beta})$，这导致了 $\boldsymbol{\beta}$ 估计值符号反向。]那么：

$$\begin{aligned}\ln f(t \mid \mathbf{x}, \boldsymbol{\beta}, \alpha) &= \ln[\exp(\mathbf{x}'\boldsymbol{\beta})\alpha t^{\alpha-1} \exp(-\exp(\mathbf{x}'\boldsymbol{\beta})t^{\alpha})] \\ &= \mathbf{x}'\boldsymbol{\beta} + \ln \alpha + (\alpha - 1)\ln t + \exp(\mathbf{x}'\boldsymbol{\beta})t^{\alpha}\end{aligned}$$

而且：

$$\begin{aligned}\ln S(t \mid \mathbf{x}, \boldsymbol{\beta}, \alpha) &= \ln[\exp(-\exp(\mathbf{x}'\boldsymbol{\beta})t^{\alpha})] \\ &= -\exp(\mathbf{x}'\boldsymbol{\beta})t^{\alpha}\end{aligned}$$

似然函数(17.17)变成:

$$\ln L=\sum_i\left[\delta_i\{\mathbf{x}_i'\boldsymbol{\beta}+\ln\alpha+(\alpha-1)\ln t_i-\exp(\mathbf{x}_i'\boldsymbol{\beta})t_i^\alpha\}-(1-\delta_i)\exp(\mathbf{x}_i'\boldsymbol{\beta})t_i^\alpha\right] \tag{17.19}$$

$\boldsymbol{\beta}$ 与 α 的一阶条件是:

$$\frac{\partial\ln L}{\partial\boldsymbol{\beta}}=\sum_i(\delta_i-\exp(\mathbf{x}_i'\boldsymbol{\beta})t_i^\alpha)\mathbf{x}_i=\mathbf{0}$$

$$\frac{\partial\ln L}{\partial\alpha}=\sum_i\delta_i(1/\alpha+\ln t_i)-\ln t_i\exp(\mathbf{x}_i'\boldsymbol{\beta})t_i^\alpha=0$$

很明显,一致性需要强假设。例如,甚至在未删失情况下,$\mathrm{E}[\partial\ln \mathrm{L}/\partial\boldsymbol{\beta}]=0$ 要求 $\mathrm{E}[T^\alpha|\mathbf{x}]=\exp(\mathbf{x}'\boldsymbol{\beta})$。

17.6.6 *模型估计值的应用*

对非线性回归模型估计值进行解释的通常方法是,考察回归元对条件均值的效应。若 $\gamma=\exp(\mathbf{x}'\boldsymbol{\beta})$,则由表 17.4 知,完整威布尔持续期限具有增值 $\mathrm{E}[T^*|\mathbf{x}]=\exp(-\mathbf{x}'\boldsymbol{\beta}/\alpha)\Gamma(\alpha^{-1}+1)=\exp(-\mathbf{x}'\boldsymbol{\beta}/\alpha)\Gamma(\alpha^{-1})/\alpha$。人们在 $\mathbf{x}$ 的各种不同值处计算完整时期的期望长度。例如,对于已知年龄、性别以及教育水平的人,就能预测出完全失业的长度。

参数回归模型除了预测样本均值外,还可预测持续期限的其他方面。例如,关注内容在于完成失业时期中居民总时间的多少份额归因于超过特定长度或者被特定社会经济群体的个体所经历。持续期限的经济计量学模型关注协变量的作用,但值得注意的是,它特别涉及风险函数的形状,这是因为某些经济理论对风险函数的形状做出了明确预测。

尽管有这些可能,但对参数持续期限模型估计值进行解释经常关注威布尔风险率 $\lambda(t)=\gamma\alpha t^{\alpha-1}$,以及它如何随时间和回归元的变动而变化。正如 17.3.2 节所提及的,当 $\alpha>1$ 时,这个风险率递增,而当 $\alpha<1$ 时,风险率递减,因此,很明显当 $\alpha=1$ 时,单侧检验成为关注焦点。就回归元的变动而言,有:

$$d\lambda(t)/d\mathbf{x}=\exp(\mathbf{x}'\boldsymbol{\beta})\alpha t^{\alpha-1}\boldsymbol{\beta}=\lambda(t)\boldsymbol{\beta}$$

所以回归元变动具有风险函数变化的乘法效应。因此,正的 β_j 系数蕴含,当 $\mathbf{x}$ 成分增大时,风险率将变大。因而,当 $\beta_j>0$ 时,x_j 增大会导致失效风险变大,从而导致期望持续期限减少。

17.6.7 *最小二乘估计*

对完全参数模型进行估计,与其用 MLE 不如用最小二乘法,这一点类似于删失 Tobit 模型。虽然在实际应用中很少看到最小二乘回归,原因在于一些方法仍依赖于对密度的正确设定,而且其有效性也不如 MLE 好,但我们仍然阐述它的一些结果。

我们以指数持续期限回归模型开始。于是，$\mathrm{E}[T|\mathbf{x}]=1/\gamma=\exp(\mathbf{x}'\boldsymbol{\beta})$，因此，$t_i$ 对 $\exp(\mathbf{x}'\boldsymbol{\beta})$ 的 NLS 回归通过 $\boldsymbol{\beta}$ 的非一致估计量给出了一致估计。否则，把指数持续期限模型写成 $\ln t=\mathbf{x}'\boldsymbol{\beta}+u$，其中，$u$ 服从极值分布（参见 17.7.2 节）。那么，$\mathrm{E}[\ln T|\mathbf{x}]=\mathbf{x}'\boldsymbol{\beta}-c$，其中 $c\simeq 0.5722$ 是一个欧拉常值。因而，借助于 $\ln t_i$ 对 $\mathbf{x}_i$ 的线性回归，能一致地估计 $\boldsymbol{\beta}$。对于右删失，我们需要获得解析删失矩，这一点对指数也是可能的。

利用基弗（Kiefer，1988，第 665 页）的更一般结果，可进一步加以推广。他考虑了满足 $\phi(\mathbf{x}'\boldsymbol{\beta})=\exp(\mathbf{x}'\boldsymbol{\beta})$ 的 PH 模型。于是，有：

$$\lambda(t|\mathbf{x})=\lambda_0(t,\alpha)\exp(\mathbf{x}'\boldsymbol{\beta})$$

那么，基准综合风险的表达式可如下推导：

$$\begin{aligned}\int_0^t \lambda(s|\mathbf{x})ds &= \int_0^t \lambda_0(s,\alpha)\exp(\mathbf{x}'\boldsymbol{\beta})ds \\ \Lambda(t|\mathbf{x}) &= \Lambda_0(t,\alpha)\exp(\mathbf{x}'\boldsymbol{\beta}) \\ \ln\Lambda(t|\mathbf{x}) &= \ln\Lambda_0(t,\alpha)+\mathbf{x}'\boldsymbol{\beta} \\ -\ln\Lambda_0(t,\alpha) &= \mathbf{x}'\boldsymbol{\beta}-\ln\Lambda(t|\mathbf{x}) \\ &= \mathbf{x}'\boldsymbol{\beta}+u\end{aligned} \tag{17.20}$$

其中，误差项 $u=-\ln\Lambda(t|\mathbf{x})$ 服从第 I 类极值分布。

不管对基准风险怎样选择，这一结果都成立。我们利用下述方法，对此结果加以解释。对于基准风险 $\lambda_0(t,\alpha)$ 的特殊选择来说，因变量 t 的一种方便变换是 $-\ln\Lambda_0(t,\alpha)$，因为它能被表述成具有服从第 I 类极值分布误差项的线性回归模型。就指数而言说，正如已经讨论的，$\ln\Lambda_0(t,\alpha)=\ln t$；而对于威布尔情况，$\ln\Lambda_0(t,\alpha)=\alpha\ln t$。在删失样本条件下，我们利用删失第 I 类极值的结果得到 $\mathrm{E}[\ln\Lambda_0(t,\alpha)|T>t^*]$，然后沿用赫克曼两步法。这些结果也可用作简单诊断学的基础；此专题将在下一章讨论。

17.7 某些重要的持续期限模型

在持续期限回归分析所使用的公式中，最广泛运用的或许是比例风险模型。不过，熟悉 17.7.2 节曾经讨论过的它的某些变形以及加速失效时间（AFT）模型也是有益的。

17.7.1 *比例风险模型*

如同前面提及的，比例风险模型（**proportional hazard model**）的条件风险率 $\lambda(t|\mathbf{x})$，可被分解成如下独立函数，

$$\lambda(t|\mathbf{x})=\lambda_0(t,\boldsymbol{\alpha})\phi(\mathbf{x},\boldsymbol{\beta}) \tag{17.21}$$

其中，$\lambda_0(t,\boldsymbol{\alpha})$ 称为基准风险（**baseline hazard**），它只是 t 的函数，而 $\phi(\mathbf{x},\boldsymbol{\beta})$ 只是 $\mathbf{x}$ 的函数。通常，$\phi(\mathbf{x},\boldsymbol{\beta})=\exp(\mathbf{x}'\boldsymbol{\beta})$。多项式基准风险在文献中颇为普遍。

所有形式为式(17.21)的风险函数 $\lambda(t|\mathbf{x})$ 都与基准风险成比例,其标度因子 $\phi(\mathbf{x},\boldsymbol{\beta})$ 不是 t 的显函数。广泛用作参数 $\boldsymbol{\beta}$ 的PH模型,在没有关于 $\lambda_0(\cdot)$ 函数形式的设定条件下能被一致地估计出来(参见17.8节)。

指数、威布尔以及冈珀茨回归模型都是PH模型,这是因为它们的风险分别是 $\exp(\mathbf{x}'\boldsymbol{\beta})$、$\exp(\mathbf{x}'\boldsymbol{\beta})\alpha t^{\alpha-1}$ 以及 $\exp(\mathbf{x}'\boldsymbol{\beta})\exp(\alpha t)$。

失业持续期限应用中,特别使用的另一个PH模型例子是分段常数风险模型(**piecewise constant hazard model**),即将 $\lambda_0(t,\boldsymbol{\alpha})$ 设成具有 k 段的阶梯函数,所以:

$$\lambda_0(t,\boldsymbol{\alpha})=e^{\alpha_j}, \quad c_{j-1}\leqslant t<c_j,\ j=1,\cdots,k \tag{17.22}$$

其中 $c_0=0$, $c_k=\infty$,其他分割点 $c_1,\cdots,c_{k-1}$ 都是设定的,而参数 $\alpha_1,\cdots,\alpha_k$ 均要加以估计。这些参数都已被指数化,以便保证 $\lambda_0(t,\boldsymbol{\alpha})>0$。与具有唯一基准风险参数的诸如威布尔模型相比,这个模型具有更多待估的基准风险参数,但利用充分大的数据集合后仍是实用的。

在不可观测异质性条件下PH模型的可识别性,将由18.3节加以讨论。

17.7.2 *加速失效时间模型*

首先,通过对 $\ln t$ 而不是 t 加以建模,得到AFT模型。回归模型是对 $\ln t$ 设定成:

$$\ln t=\mathbf{x}'\boldsymbol{\beta}+u \tag{17.23}$$

并且 u 的各种不同分布会产生不同的AFT模型。由于 $\ln t$ 取值为 $(-\infty,\infty)$,所以,u 的分布可以是 $(-\infty,\infty)$ 上的任何连续分布。

加速失效时间(**accelerated failure time**)的产生,是因为 $t=\exp(\mathbf{x}'\boldsymbol{\beta})v$,其中,$v=e^u$,具有风险率 $\lambda(t|\mathbf{x})=\lambda_0(v)\exp(\mathbf{x}'\boldsymbol{\beta})$,这里,基准风险 $\lambda_0(v)$ 并不依赖于 t。将 $v=t\exp(-\mathbf{x}'\boldsymbol{\beta})$ 代入,得到风险:

$$\lambda(t|\mathbf{x})=\lambda_0(t\exp(-\mathbf{x}'\boldsymbol{\beta}))\exp(-\mathbf{x}'\boldsymbol{\beta}) \tag{17.24}$$

当 $\exp(-\mathbf{x}'\boldsymbol{\beta})>1$ 时,这是基准风险 $\lambda_0(t)$ 的加速式,而当 $\exp(-\mathbf{x}'\boldsymbol{\beta})<1$ 时,则是基准风险 $\lambda_0(t)$ 的减速式。

若 $u\sim\mathcal{N}[0,\sigma^2]$,则得到 t 的对数正态模型;当将 u 设定成逻辑斯蒂分布,则得出对数逻辑斯蒂模型。通过令 u 具有密度 $f(u)=\exp(\alpha u-e^u)/\Gamma(\alpha)$,也可获得伽玛模型作为AFT模型。

威布尔模型与指数模型是唯一既是PH形式又是AFT形式的模型。后者通过令 u 是 αw 而得到,其中,w 服从密度为 $f(w)=e^w\exp(-e^w)$ 的极值分布。

另一些持续期限模型,可通过考察 $g(t)=\mathbf{x}'\boldsymbol{\beta}+u$,即把上式看成一个变换而不是 $g(t)=\ln t$ 而得到,这是变换类型模型的一个成员,例如,该变换类型模型包括了Box-Cox回归模型。

17.7.3 灵活风险函数

一些模型与其以设定 pdf 开始，不如从设定风险率开始。例如，把风险率设定成 t 的二次型，比如 $\lambda(t)=\mathbf{x}'\boldsymbol{\beta}+a_1t+a_2t^2$。这就出现 U 形状风险函数。相应的综合风险是 $\Lambda(t)=(\mathbf{x}'\boldsymbol{\beta})t+(a_1/2)t^2+(a_2/3)t^3$。已知 $\lambda(t)$ 与 $\Lambda(t)$ 时，能利用前面结果直接构建其对数似然函数。

这种方法的缺陷是，可能出现 λ 与 Λ 的负值，同时相应 pdf 积分可能不一定为 1，从而导致风险率有缺陷。

17.8 考克斯 PH 模型

对于单时期持续期限数据来说，完全参数模型在删失情况下相对容易直接估计，但倘若参数模型的任何部分被错误设定，就产生非一致参数估计。解决这个问题的一种方法是，选取灵活的参数函数形式，从而对错误设定提供某种防范。原则上讲，这是一个有效方法，但此类灵活函数形式的识别与估计并不总是简单易行。一个例子是广义伽玛模型，许多使用者发现，很难对其进行估计。

幸运的是，半参数方法并不需要对分布完全设定，这种方法与针对 Tobit 模型所提出的半参数方法有相当大的差异，因为这种方法建立在风险率模型基础上，而这样的风险率模型在 Tobit 情况下没有什么有意义的科学解释，类似于删失情况下导致模型出现稳健性问题。另外，与 Tobit 情况不同，从经验上讲，半参数方法被认为是成功的，以致它已经成为生存数据的标准方法。

17.8.1 比例风险模型

研究起点是，提出一种特殊的风险率函数形式，即由 17.7.1 节引入的比例风险模型，条件风险率 $\lambda(t|\mathbf{x})$ 被因式分解为

$$\lambda(t|\mathbf{x},\boldsymbol{\beta})=\lambda_0(t)\phi(\mathbf{x},\boldsymbol{\beta}) \tag{17.25}$$

的独立函数。如上所述，函数 $\lambda_0(t)$ 被称为基准风险，并仅仅是 t 的函数。函数 $\phi(\mathbf{x},\boldsymbol{\beta})$ 仍然只是 $\mathbf{x}$ 的函数，最初我们考虑时不变回归元 $\mathbf{x}$ 的情况，而稍后则放松这个假设。还要考虑半参数模型，那里 $\lambda_0(t)$ 的函数形式未加以设定，而对 $\phi(\mathbf{x},\boldsymbol{\beta})$ 的函数形式则完全设定。

对 $\phi(\mathbf{x},\boldsymbol{\beta})$ 的一种最普遍选取是指数形式：

$$\phi(\mathbf{x},\boldsymbol{\beta})=\exp(\mathbf{x}'\boldsymbol{\beta}) \tag{17.26}$$

这不仅确保 $\phi(\mathbf{x},\boldsymbol{\beta})>0$，还使得对系数很容易地给予解释。假定第 j 个回归元 x_j 增加一个单位，同时其他回归元保持不变，则：

$$\begin{aligned}\lambda(t|\mathbf{x}_{new},\boldsymbol{\beta})&=\lambda_0(t)\exp(\mathbf{x}'\boldsymbol{\beta}+\beta_j)\\&=\exp(\beta_j)\lambda(t|\mathbf{x},\boldsymbol{\beta})\end{aligned} \tag{17.27}$$

因而，新的风险率是 $\exp(\beta_j)$ 乘以原来风险率，而其风险变化是 $1-\exp(\beta_j)$ 乘以原

来风险率。不过，当人们使用微分法，则风险变化是 β_j 乘以原来风险率，这是因为：

$$\partial\lambda(t|\mathbf{x},\boldsymbol{\beta})/\partial x_j=\lambda_0(t)\exp(\mathbf{x}'\boldsymbol{\beta})\beta_j=\beta_j\lambda(t|\mathbf{x},\boldsymbol{\beta}) \tag{17.28}$$

这与非微分法结果相一致，因为 $\exp(\beta_j)\simeq 1+\beta_j$。统计软件经常报告出既有 β_j 又有 $\exp(\beta_j)$ 的估计值及其相关的置信区间。

对于 $\phi(\mathbf{x},\boldsymbol{\beta})$ 的更一般形式，回归元变化也能被解释成对原来风险具有乘法效应，这是因为：

$$\begin{aligned}\partial\lambda(t|\mathbf{x},\boldsymbol{\beta})/\partial\mathbf{x}&=\lambda_0(t)\partial\phi(\mathbf{x},\boldsymbol{\beta})/\partial x_j\\&=\lambda(t|\mathbf{x},\boldsymbol{\beta})\times[\partial\phi(\mathbf{x},\boldsymbol{\beta})/\partial x_j]/\phi(\mathbf{x},\boldsymbol{\beta})\end{aligned} \tag{17.29}$$

这需要 $\boldsymbol{\beta}$ 的知识，但并不要求基准风险 $\lambda_0(t)$ 的知识。

一个重要问题是对 PH 模型加以识别。这将由下一章在更一般背景下进行讨论，那里考虑模型有不可观测异质性的问题。

17.8.2 *偏似然估计*

考克斯(Cox, 1972, 1975)曾经提出一种估计 PH 模型中 $\boldsymbol{\beta}$ 的方法，那里并没有要求同时估计基准风险函数 $\lambda_0(t)$，倘若令人满意的基准风险估计能在估计 $\boldsymbol{\beta}$ 之后重新得到。这里所阐述的结果，适合于独立删失数据与结数据。

设置类似于 17.5 节，对失效数据加以排序，并将观测值分成哪些是停止的，或哪些是处于风险之中的失效时间。设 $t_1<t_2<\cdots<t_j<\cdots<t_k$ 表示样本量为 N 的样本中观测到时期的离散失效时间，$N\geqslant k$。风险集合 $R(t_j)$ 被定义成刚好在第 j 个有序失效时间之前处于失效的个体集合，$D(t_j)$ 是在时间 t_j 停止的实验者集合，而 d_j 表示在时间 t_j 停止的数量。概括地讲，我们有：

$$\begin{aligned}R(t_j)&=\{l:\ t_l\geqslant t_j\}=\text{在 } t_j \text{ 时处于风险的时期集}\\D(t_j)&=\{l:\ t_l=t_j\}=\text{在 } t_j \text{ 时完整时期集合}\\d_j&=\textstyle\sum_l\mathbf{1}(t_l=t_j)=\text{在 } t_j \text{ 时完整时期数量}\end{aligned} \tag{17.30}$$

在时间 t_j 时风险集合包括尚未完成的或尚未删失的所有时期。在 $d_j>1$ 的情况下，可能有结数据(**tied data**)。

现在，考察在时间 t_j 特定风险时期将结束的概率。时期 j 结束的真实时期概率，等于时期 j 失效的条件概率被风险集合 $R(t_j)$ 中任何个体时期失效的条件概率去除。后者概率是 $R(t_j)$ 中每一个个体失效的条件概率之和。于是有：

$$\begin{aligned}\Pr[T_j=t_j|R(t_j)]&=\frac{\Pr[T_j=t_j|T_j\geqslant t_j]}{\sum_{l\in R(t_j)}\Pr[T_l=t_l|T_l\geqslant t_j]}\\&=\frac{\lambda_j(t_j|\mathbf{x}_j,\boldsymbol{\beta})}{\sum_{l\in R(t_j)}\lambda_l(t_j|\mathbf{x}_l,\boldsymbol{\beta})}\\&=\frac{\phi(\mathbf{x}_j,\boldsymbol{\beta})}{\sum_{l\in R(t_j)}\phi(\mathbf{x}_l,\boldsymbol{\beta})}\end{aligned}$$

其中，最后一行基准风险因子 $\lambda_0(t_j)$ 被省略了，这是因为 PH 假设的缘故。(因而，这一模型中的截距是不可识别的。)上述基准风险能被省略的结果提供了估计 $\boldsymbol{\beta}$ 的

基础。然而，我们必须控制出现持续期限分组时可能发生的结持续期限。

当对持续期限进行分组时，更有可能产生结。若数据包括结（也就是说，在特定时间有一个以上失效发生），则需要加以调整。假如在时间 t_j 处有两个结，个体 j_1 与 j_2 具有回归元 x_{j1} 与 x_{j2}。如果 j_1 在 j_2 之前失效，那么其概率是：

$$\phi(\mathbf{x}_{j1},\boldsymbol{\beta})\Big/\sum_{l\in R(t_j)}\phi(\mathbf{x}_l,\boldsymbol{\beta})+\phi(\mathbf{x}_{j2},\boldsymbol{\beta})\Big/\sum_{l\in R(t_j)}\phi(\mathbf{x}_l,\boldsymbol{\beta})$$

其中，当实验者 j_1 被排除时，$R_1(t_j)$等于 $R(t_j)$。当 j_2 在 j_1 之前失效时，会产生类似项，而似然贡献则是这两种可能性之和。一旦出现许多结，准确似然变得相当复杂。

归因于布雷斯洛（Breslow）和皮托（Peto）的标准近似，参见考克斯和奥克斯（Cox and Oakes，1984），设：

$$\Pr[T_j=t_j\,|\,j\in R(t_j)]\simeq\frac{\prod_{m\in D(t_j)}\phi(\mathbf{x}_m,\boldsymbol{\beta})}{[\sum_{l\in R(t_j)}\phi(\mathbf{x}_l,\boldsymbol{\beta})]^{d_j}}\tag{17.31}$$

其中，$D(t_j)$表示在时间 t_j 处死亡的实验者集合，而 d_j 表示在时间 t_j 处死亡的数量。倘若在时间 t_j 处失效的数量相对于风险数量而言很小，则这种近似会表现很好。

考克斯把偏似然函数定义成 k 个有序失效时间上的联合乘积 $\Pr[T_j=t_j\,|\,j\in R(t_j)]$。于是，有：

$$\mathrm{L_p}(\boldsymbol{\beta})=\prod_{j=1}^{k}\frac{\prod_{m\in D(t_j)}\phi(\mathbf{x}_m,\boldsymbol{\beta})}{[\sum_{l\in R(t_j)}\phi(\mathbf{x}_l,\boldsymbol{\beta})]^{d_j}}\tag{17.32}$$

考克斯提出通过极小化对数偏似然函数

$$\ln\mathrm{L_p}=\sum_{j=1}^{k}\Big[\sum_{m\in D(t_j)}\ln\phi(\mathbf{x}_m,\boldsymbol{\beta})-d_j\ln\Big(\sum_{l\in R(t_j)}\phi(\mathbf{x}_l,\boldsymbol{\beta})\Big)\Big]\tag{17.33}$$

来估计 $\boldsymbol{\beta}$。删失时期仅仅出现在 $\ln\mathrm{L_p}$ 中的第二项，这是因为对观测到的死亡来说它们没有贡献，一直到它们被删失为止，都会影响到风险集合的大小。式(17.33)被重新写成：

$$\ln\mathrm{L_p}(\boldsymbol{\beta})=\sum_{i=1}^{N}\delta_i\Big[\ln\phi(\mathbf{x}_i,\boldsymbol{\beta})-\ln\Big(\sum_{l\in R(t_i)}\phi(\mathbf{x}_l,\boldsymbol{\beta})\Big)\Big]\tag{17.34}$$

其中，对于未删失观测值指示变量有 $\delta_i=1$，否则有 $\delta_i=0$。

就通常设定 $\phi(\mathbf{x},\boldsymbol{\beta})=\exp(\mathbf{x}'\boldsymbol{\beta})$ 而言，有 $\ln\phi(\mathbf{x},\boldsymbol{\beta})=\mathbf{x}'\boldsymbol{\beta}$，故得到的一阶条件为：

$$\frac{\partial\ln\mathrm{L_p}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}=\sum_{i=1}^{N}\delta_i[\mathbf{x}_i-\mathbf{x}_i^*(\boldsymbol{\beta})]=\mathbf{0}$$

其中，$\mathbf{x}_i^*(\boldsymbol{\beta})=\sum_{l\in R(t_i)}\mathbf{x}_l\exp(\mathbf{x}_l'\boldsymbol{\beta})/\sum_{l\in R(t_i)}\exp(\mathbf{x}_l'\boldsymbol{\beta})$表示在失效时间 t_i 处位于风险中的实验者的回归元 $\mathbf{x}_l$ 的加权平均值。

偏似然是有限信息似然，这是因为基准风险 $\lambda_0(t)$ 被省略了，它既不是条件似然，也不是边缘似然。统计学文献中，对 $L_p(\boldsymbol{\beta})$ 是否是有效似然函数进行了大量讨论。可以证明[安德森等人(Andersen, et al., 1993)]，尽管 $\ln L_p$ 不是完整函数，但求 $\ln L_p$ 极大值的估计量 $\boldsymbol{\beta}$ 却是一致的。也可参见卡尔布弗莱舍和普伦蒂斯(Kalbfleisch and Prentice, 2002，第 91～101 页)以及兰开斯特(Lancaster, 1990，第 9 章)。

借助于类似 ML 情况，$\mathbf{A}(\boldsymbol{\beta})=-\mathbf{B}(\boldsymbol{\beta})$ 进行简化，可应用第 5 章极值估计的结果，因此有：

$$\hat{\boldsymbol{\beta}} \overset{a}{\sim} \mathcal{N}\left[\boldsymbol{\beta},\ \left(-\mathrm{E}\left[\frac{\partial^2 \ln L_p(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right]\right)^{-1}\right] \tag{17.35}$$

虽然针对完全参数 PH 模型诸如威布尔模型，把 MLE 与偏似然估计量加以比较，揭示出损失了相对很小的有效性，但估计量却是无效的。

17.8.3 *考克斯 PH 模型的生存函数*

许多研究都停留在对 $\boldsymbol{\beta}$ 的估计上，内容涉及利用式(17.28)或式(17.29)测量回归元变动对基准风险的影响。而另一些研究则对基准风险函数的形状感兴趣。对于 PH 模型，通过求偏似然极大值得到 $\boldsymbol{\beta}$，就可能得到基准风险函数或生存函数的非参数估计。这种估计类似于 17.5.1 节中的卡普兰—迈耶估计量。

利用 $S(t\,|\,\mathbf{x},\boldsymbol{\beta})=\exp\left[-\int_0^t \lambda_0(s)\phi(\mathbf{x},\boldsymbol{\beta})ds\right]$，并定义 $S_0(t)=\exp\left[-\int_0^t \lambda(s)ds\right]$。我们得到与 PH 风险函数有关的生存函数：

$$S(t\,|\,\mathbf{x},\boldsymbol{\beta})=S_0(t)^{\phi(\mathbf{x},\boldsymbol{\beta})}$$

现在，假定离散时间公式在离散失效时间 t_j 处具有基准风险率 $1-\alpha_j, j=1,\cdots,k$。利用由下一节给出的某些大量代数运算，得出估计 $\hat{\alpha}_j$，$\hat{\alpha}_j$ 是

$$\sum_{l\in D(t_j)}^{k} \frac{\phi(\mathbf{x}_l,\hat{\boldsymbol{\beta}})}{1-\hat{\alpha}_j^{\phi(\mathbf{x}_l,\hat{\boldsymbol{\beta}})}} = \sum_{m\in R(t_j)} \phi(\mathbf{x}_m,\hat{\boldsymbol{\beta}}),\quad j=1,\cdots,k \tag{17.36}$$

的解，其中 $\hat{\boldsymbol{\beta}}$ 是 $\boldsymbol{\beta}$ 的偏似然估计量，$D(t_j)$ 表示在时间 t_j 处死亡的实验者，而 $R(t_j)$ 表示在时间 t_j 处位于风险之中的实验者。由 17.3.3 节对离散时间风险的讨论知道，基准生存数 $S_0(t)=\prod_{j|t_j\leqslant t}\alpha_j$，即瞬时条件生存概率的累积乘积。于是，估计基准生存函数是：

$$\hat{S}_0(t)=\prod_{j|t_j\leqslant t}\hat{\alpha}_j \tag{17.37}$$

若不存在回归元，则 $\hat{S}_0(t_0)$ 简化成卡普兰—迈耶估计量，即正规化 $\phi(\mathbf{x}_l,\boldsymbol{\beta})=1$，同时表达式得出风险率 $1-\hat{\alpha}_j=d_j/r_j$。倘若有回归元但没有结存在，则由表达式得出，风险率 $1-\hat{\alpha}_j=\phi(\mathbf{x}_j,\hat{\boldsymbol{\beta}})/\sum_{m\in R(t_j)}\phi(\mathbf{x}_j,\hat{\boldsymbol{\beta}})$。

对于具有回归元 $\mathbf{x}=\mathbf{x}^*$ 的个体来说，生存函数可利用

$$\hat{S}(t\,|\,\mathbf{x}^*,\boldsymbol{\beta})=\hat{S}_0(t)^{\phi(\mathbf{x}^*,\hat{\boldsymbol{\beta}})}$$

加以估计。回归元的线性变换并不会改变 $\boldsymbol{\beta}$ 的估计值，只是线性变换会改变基准风险函数。例如：

$$\begin{aligned}\lambda(t|\mathbf{x},\boldsymbol{\beta})&=\lambda_0(t)\exp(\mathbf{x}'\boldsymbol{\beta})\\&=\lambda_0(t)\exp(\bar{\mathbf{x}}'\boldsymbol{\beta})\exp((\mathbf{x}-\bar{\mathbf{x}})'\boldsymbol{\beta})\\&=\lambda_0^*(t)\exp((\mathbf{x}-\bar{\mathbf{x}})'\boldsymbol{\beta})\end{aligned}$$

其中，新的基准风险是 $\lambda_0^*(t\exp((\mathbf{x}-\bar{\mathbf{x}})'\boldsymbol{\beta})$ 。因此，对每个回归元都减去样本均值，将改变基准风险，故在解释基准风险函数或生存函数时，需要小心谨慎。

另外，尽管估计基准风险对于计算和比较特定个体分组来说是有用的，但它可能表现出非常的不连贯特性，为了令解释容易理解，可对它们进行某种光滑处理。

17.8.4 *生存函数的推导*

沿着卡尔布弗莱舍和普伦蒂斯（Kalbfleisch and Prentice, 2002，第 114～118 页）线索，我们得到由式（17.36）给出的 α_j 的估计方程。

持续期限为 t_j 的实验者具有如下似然贡献，即生存时间 $t>t_{j-1}$ 的概率减去生存时间 $t>t_j$ 的概率。这就是：

$$\begin{aligned}S(t_j|\mathbf{x},\boldsymbol{\beta})-S(t_{j+1}|\mathbf{x},\boldsymbol{\beta})&=S_0(t_j)^{\phi(\mathbf{x},\boldsymbol{\beta})}-S_0(t_{j+1})^{\phi(\mathbf{x},\boldsymbol{\beta})}\\&=(\alpha_j^{-1}S_0(t_{j+1}))^{\phi(\mathbf{x},\boldsymbol{\beta})}-S_0(t_{j+1})^{\phi(\mathbf{x},\boldsymbol{\beta})}\\&=(\alpha_j^{-\phi(\mathbf{x},\boldsymbol{\beta})}-1)S_0(t_{j+1})^{\phi(\mathbf{x},\boldsymbol{\beta})}\end{aligned}$$

这里，用到了 $S_0(t_{j+1})=\Pi_{l=1}^{j}\alpha_l=\alpha_jS_0(t_j)$。

对于在时间 t_j 被删失的那些实验者来说，其似然贡献是生存 $t>t_j$ 的概率，或者 $S_0(t_{j+1})^{\phi(\mathbf{x},\boldsymbol{\beta})}$。因此，在$[t_j,t_{j-1})$内要么死亡要么被删失的实验者贡献概率 $S_0(t_{j+1})^{\phi(\mathbf{x},\boldsymbol{\beta})}=\Pi_{l=1}^{j}\alpha_l^{\phi(\mathbf{x},\boldsymbol{\beta})}$，对于死亡实验者，具有额外乘子$(\alpha_l^{\phi(\mathbf{x},\boldsymbol{\beta})}-1)$。于是，在全部失效时间上，似然函数是：

$$\mathrm{L}(\boldsymbol{\alpha},\boldsymbol{\beta})=\prod_{j=1}^{k}\Big[\prod_{l\in D(t_j)}(\alpha_j^{-\phi(\mathbf{x}_l,\boldsymbol{\beta})}-1)\prod_{m\in R(t_j)}\alpha_j^{-\phi(\mathbf{x}_m,\boldsymbol{\beta})}\Big]$$

对数似然函数是：

$$\ln\mathrm{L}(\boldsymbol{\alpha},\boldsymbol{\beta})=\sum_{j=1}^{k}\Big[\sum_{l\in D(t_j)}\ln(\alpha_j^{-\phi(\mathbf{x}_l,\boldsymbol{\beta})}-1)+\sum_{m\in R(t_j)}-\phi(\mathbf{x}_m,\boldsymbol{\beta})\ln\alpha_j\Big]$$

从而$\partial\ln\mathrm{L}(\boldsymbol{\alpha},\hat{\boldsymbol{\beta}})/\partial\alpha_j=0$ 能重新写成式（17.36）。

17.9 时变回归元

上述结果局限于回归元为如下变量的模型：比如性别回归元变量，对不同个体来说是变化的，但对已知个体来说则不随时间变化。在其他标准的横截面模型诸如 logit 模型与 Tobit 模型中，这是一种标准情形。不过，对于生存数据，一些个体在时期的几个阶段上都可以被观测到，而有关回归元在某个时间中可能取不同的

一些值。例如，在医学生存研究中，处方剂量水平对已知个体可能随时间而变化。在失业时期期间，失业救济金或许以离散方式变动。在寻找工作时，某人的婚姻状态可能变化。

使用时变协变量可能会产生两类问题。第一类问题，很明显是将时变协变量错误设定成固定变量。协变量在某个时期上的整个历史是相关的，需要考虑的事可能要求我们将某些回归元的滞后值合并到风险率的决定因素之中。第二类问题，时变协变量可能表现出反馈特性，因此，可能不是严格外生的，在持续期限模型中人们经常做出如此假定。例如，失业时期的持续期限可能依赖于个体寻找工作策略，而后者当失业持续期限延长时可能会改变。第二个例子是，治疗的药剂量水平会随病人的病情变坏或改善而变化。确定性时间变化很容易加以处理，因此，标准分析仅仅考虑到上述两类问题中的第一类，需要做出协变量是弱外生的假设；也就是说，无论支撑时间变化的过程是随机的还是确定的，我们都不需要考虑估计风险模型时所处理的参数。一些作者[比如，卡尔布弗莱舍和普伦蒂斯（Kalbfleish and Prentice，2002，第 196～200 页）]将这类时间变化称为外部的。于是，把内生时变协变量称为内部的。

特别地，当某软件包不能处理时变协变量时，一种相当简单的求解方法是，时变协变量的时期平均值代替时变协变量。不过，好的软件包都会给出较大灵活性。

考察某个个体从最初持续到时间 T 的失业时期，在此期间可以观测到转换为就业状态。设 $0<t_1<t_2<t_3<T$，其中 t_1、t_2 以及 t_3 都是此时期的中间点。假定有两个协变量 x_1 与 $x_2(t)$，它们分别是时不变的与时变的。为了简单起见，假定 x_1 为二值的，而 x_2 在阶梯形区间$[0,t_1)$、$[t_1,t_2)$、$[t_2,T)$分别取值为 $x_2(t_1)$、$x_2(t_2)$、$x_3(t_3)$。而且，假定时变回归元是外生的或时间变化形式是确定的。那么，就这个特定时期而言，能将数据写成如下三行记录，而不是一行记录：

持续期限时间

观测值	持续期限	x_1	$x_2(t)$	删失指示变量
1	t_1	1	$x_2(t_1)$	0
1	t_2	1	$x_2(t_2)$	0
1	T	1	$x_2(T)$	1

对于这种信息的解释是，我们可将观测到的全部持续期限划分成三个时段。在第一个与第二个时段期间，协变量值分别是$(1,\ x_2(t_1))$与$(1,\ x_2(t_2))$，并且没有观测到就业转换（因此，删失指示变量为 0），然后在第三个时段期间，协变量值是$(1,\ x_2(T))$，并观测到就业转换。这类似于拥有三个观测值情形，其中两个持续期限被删失了，而第三个持续期限是完整的。

现在，假定将 $x_2(t)$的当前值与滞后一期值作为适当的协变量。也就是说，在某时点上的风险率可能依赖于协变量较早时期的变化。于是，将此类数据写成如下形式：

持续期限时间

观测值	持续期限	x_1	$x_2(t)$	$x_2(t-1)$	删失指示变量
1	t_1	1	$x_2(t_1)$	0	0
1	t_2	1	$x_2(t_2)$	$x_2(t_1)$	0
1	T	1	$x_2(T)$	$x_2(t_2)$	1

这里，我们假定先于时期开始的 $x_2(t)$之值为 0。注意到，在这两个例子中，协变量 $x_2(t)$在离散时间点上变化。

尽管人们在数据集合中拥有多行元素，但如果软件以将各个元素处理成为各不相同的观测值而结束，那么这种大数据集合显得冗长并存在潜在混淆。幸运的是，计算机软件经常允许用户将时变协变量识别成回归模型定义的一部分。对于该时期已流逝的期限，人们能给出阶梯函数或连续函数。

17.9.1 *推广考克斯模型*

容易将 17.8 节的考克斯模型固定回归元分析推广到时变回归元上。

一般来说，风险函数依赖于回归元 $\mathbf{x}(t)$ 的完整时间路径，因而有：

$$\lambda(t|\mathbf{x}(t))=\lim_{\Delta t\to 0}\frac{\Pr[t\leqslant T<t+\Delta t|\mathbf{x}(t),T\geqslant t]}{\Delta t}$$

我们考察 PH 形式：

$$\lambda(t|\mathbf{x}(t))=\lambda_0(t,\boldsymbol{\alpha})\phi(\mathbf{x}(t),\boldsymbol{\beta})$$

这里做出了如下约束：只有协变量 $\mathbf{x}(t)$的当前值而不是 $\mathbf{x}(t)$的整个历史值才会起作用。

很明显，由 17.8.2 节的考克斯偏似然方法知，在每个失效时间 t_j 处，起作用的是风险集合 $R(t_j)$中那些观测值的回归元 $\mathbf{x}(t_j)$之值。因而，对于第 i 个实验者，用 $\mathbf{x}_j(t_j)$代替 $\mathbf{x}_j$。偏似值函数有类似变化，并且：

$$\ln \mathrm{L_p}=\sum_{j=1}^{k}\Big[\sum_{m\in D(t_j)}\ln\phi(\mathbf{x}_m(t_j),\boldsymbol{\beta})-d_j\ln\Big(\sum_{l\in R(t_j)}\phi(\mathbf{x}_l(t_j),\boldsymbol{\beta})\Big)\Big]$$

注意到，数据形式现在更为复杂，因为每个实验者都具有多重观测值。例如，假定时间取离散整数值，只有一个回归元，而且观测值具有完整持续期限 25，回归元为 x_1，它在[0, 5]上取值为 50，在[6, 15]上取值为 100，而在[16, 25]上取值为 200。于是，$x_1(t_1)=50$，$x_1(t_2)=100$，$x_1(t_3)=100$，$x_1(t_4)=200$，$x_1(t_5)=200$。

17.10 离散时间比例风险

当失效时间以加总时间区间比如周或月份的形式，被人们观测到或记录时，使用分组持续模型就更为恰当。

一种简单方法是建立面板数据，并对每个时期中个体失效概率的叠放 logit 或 probit 模型加以估计，其中，每个时期都具有各自截距。17.10.3 节将阐述这一内

容。不过,我们首先阐述连续时间 PH 模型的离散时间变形,参见布莱克、伦德和蒂默曼(Blake, Lunde, and Timmermann, 1999)。

17.10.1 离散时间比例风险

对于带有分组点 t_a,$a=1,\cdots,A$ 的分组数据,其离散时间风险函数定义为:

$$\lambda^d(t_a|\mathbf{x})=\Pr[t_{a-1}\leqslant T<t_a|T\geqslant t_{a-1},\mathbf{x}(t_{a-1})],\quad a=1,\cdots,A$$

允许出现时间回归元。相应的离散时间生存函数是:

$$S^d(t_a|\mathbf{x})=\Pr[T\geqslant t_{a-1}|\mathbf{x}]=\prod_{s=1}^{a-1}(1-\lambda^d(t_s|\mathbf{x}(t_s)))$$

首先,我们得到离散时间与连续时间风险之间的一般关系。离散时间风险是 $[t_{a-1},t_a)$ 中失效概率除以至少生存到时间 t_{a-1} 时的概率,所以能重新写成:

$$\lambda^d(t_a|\mathbf{x})=\frac{S(t_{a-1}|\mathbf{x})-S(t_a|\mathbf{x})}{S(t_{a-1}|\mathbf{x})}\tag{17.38}$$

其中,$S(t|\mathbf{x})$ 表示生存函数。在连续情况下,$S(t|\mathbf{x})=\exp\left(-\int_0^t\lambda(s)ds\right)$,经过一些代数运算之后,式(17.38) 变成:

$$\lambda^d(t_a|\mathbf{x})=1-\exp\left(-\int_{t_{a-1}}^{t_a}\lambda(s)ds\right)\tag{17.39}$$

现在,对于 $[t_{a-1},t_a)$ 中的 t,列举与连续 PH 模型:

$$\lambda(t)=\lambda_0(t)\exp(\mathbf{x}(t_{a-1})'\boldsymbol{\beta})$$

有关的离散时间风险。注意到,该区间内的回归元为常值,但对不同区间来说却是变化的,而 $\lambda_0(t)$ 在区间内会变动。于是,式(17.39)变成:

$$\begin{aligned}\lambda^d(t_a|\mathbf{x})&=1-\exp(-\exp(\mathbf{x}(t_{a-1})'\boldsymbol{\beta})\times\int_{t_{a-1}}^{t_a}\lambda_0(s)ds\\&=1-\exp(-\lambda_{0a}\exp(\mathbf{x}(t_{a-1})'\boldsymbol{\beta}))\\&=1-\exp(-\exp(\ln\lambda_{0a}+\mathbf{x}(t_{a-1})'\boldsymbol{\beta}))\end{aligned}\tag{17.40}$$

其中,$\lambda_{0a}=\int_{t_{a-1}}^{t_a}\lambda_0(s)ds$。有关的离散时间生存函数是:

$$S^d(t_a|\mathbf{x})=\prod_{s=1}^{a-1}\exp(-\exp(\ln\lambda_{0s}+\mathbf{x}(t_{s-1})'\boldsymbol{\beta}))\tag{17.41}$$

第 i 个实验者的密度是,存活实验者每个时期的生存函数的乘积乘以失效时的风险。由式(17.40)与式(17.41)可得,似然函数是:

$$\begin{aligned}L(\boldsymbol{\beta},\lambda_{01},\cdots,\lambda_{0A})=&\prod_{i=1}^{N}\left[\prod_{s=1}^{a_i-1}\exp(-\exp(\ln\lambda_{0s}+\mathbf{x}_i(t_{s-1})'\boldsymbol{\beta}))\right]\\&\times(1-\exp(-\exp(\ln\lambda_{0a_i}+\mathbf{x}_i(t_{a-1})'\boldsymbol{\beta})))\end{aligned}\tag{17.42}$$

其中,为了简单起见,忽略了删失,并假定第 i 个实验者失效发生在时间 t_{a_i},假定至少有一个失效发生在每个区间 $[t_{a-1},t_a)$ 之内。

MLE 对式(17.42)求关于 $\boldsymbol{\beta}$ 与 $\lambda_{01},\cdots,\lambda_{0A}$ 的极小值。在特殊情况下，偏似然渐近地等价于 MLE，尽管它们各不相同。更为简洁的模型要对 $\lambda_{01},\cdots,\lambda_{0A}$ 施加某种结构，比如关于时间为多项式的。甚至用完全参数模型，诸如威布尔模型作为更重要的结构，即设 $\lambda_{0s}=\int_{t_{a-1}}^{t_a}\alpha s^{\alpha-1}ds$。

17.10.2 哈恩和豪斯曼方法

哈恩和豪斯曼(Han and Hausman, 1990)曾经提出一种灵活方法，重新获得相对容易实施的基准风险，他们的这项研究工作早于布莱克等人(Blake et al., 1999)，却类似于迈耶(Meyer, 1990)与末吉(Sueyoshi, 1992)。尽管保持协变量函数的参数形式[比如 $\exp(\mathbf{x}'\boldsymbol{\beta})$]，却考虑到了基准风险 $\lambda_0^d(t)$ 设定中的相当灵活性。而且，它具有明显处理离散持续期限数据以及提供更容易适应的离散数据的特性，诸如结观测值与不可观测异质性的框架。对离散数据来说，结观测值是一个重要问题；例如，就失业持续期限而言，许多失业时期的终止，可能与失业救济金时期的结束相一致(通常，在美国为 26 周)。

讨论起点是第 i 个观测值的风险率 $\lambda_i(\tau)$，该风险率表示时期终止于区间 $(\tau,\tau+\Delta)$ 之内的条件概率，以 PH 形式写成：

$$\lambda_i(\tau)=\lambda_0(\tau)\exp(-\mathbf{x}_i'\boldsymbol{\beta})$$

其中，$\lambda_0(\tau)$ 表示基准风险。于是如同式(17.20)所证明的，在积分之后取对数，然后重新整理得出：

$$\Lambda_0(\tau)-\mathbf{x}_i'\boldsymbol{\beta}=\varepsilon_i \tag{17.43}$$

其中，$\Lambda_0(t)=\ln\int_0^t\lambda_0(\tau)d\tau$ 表示综合基准风险的对数，而 $\varepsilon_i=\ln\int_0^t\lambda_i(\tau)d\tau$。从而，概率是：

$$\Pr[\text{在时期失效}]=\int_{\Lambda_0(t-1)-\mathbf{x}_i'\boldsymbol{\beta}}^{\Lambda_0(t)-\mathbf{x}_i'\boldsymbol{\beta}}f(\varepsilon)d\varepsilon$$

当第 i 个人在时期 t 经历失效时令 $y_{it}=1$，否则令 $y_{it}=0$。那么，N 个观测值的联合似然由

$$\ln\mathrm{L}(\boldsymbol{\beta},\Lambda_0(1),\cdots,\Lambda_0(T))=\sum_{i=1}^{N}\sum_{t=1}^{T}y_{it}\ln\left[\int_{\Lambda_0(t-1)-\mathbf{x}_i'\boldsymbol{\beta}}^{\Lambda_0(t)-\mathbf{x}_i'\boldsymbol{\beta}}f(\varepsilon)d\varepsilon\right] \tag{17.44}$$

给出，而基准风险参数($\Lambda_0(1),\cdots,\Lambda_0(T)$)与 $\boldsymbol{\beta}$ 都是以一种灵活方式(也就是说，在没有施加特定函数形式的条件下)加以估计。

当然，对数似然的积分是 cdf[$\Lambda_0(t-1)-\mathbf{x}_i'\boldsymbol{\beta}$, $\Lambda_0(t)-\mathbf{x}_i'\boldsymbol{\beta}$] 的差。这种表达式的精确形式依赖于 cdf 的函数形式。若假定随机误差 ε_i 服从标准正态分布，则对数似然取有序 probit 形式；在极值分布假设下，对数似然取有序 logit 形式。具体地讲，在正态性下，第 i 项的积分形式为：

$$\Pr[\Lambda_0(t)<\mathbf{x}_i'\boldsymbol{\beta}+\varepsilon_i\leqslant\Lambda_0(t+1)]=\Phi(\Lambda_0(t+1)-\mathbf{x}_i'\boldsymbol{\beta})-\Phi(\Lambda_0(t)-\mathbf{x}_i'\boldsymbol{\beta})$$

与偏似然方法——把基准风险处理成冗余函数并且剔除它——相比,哈恩和豪斯曼(Han and Hausman, 1990)方法则是,以适度的计算成本估计出所有未知参数。他们的蒙特卡罗结果表明,该方法灵活,并且能很好地逼近任何风险函数,并且不要求强的函数形式假设。

17.10.3 *离散时间二值选择*

离散持续期限数据的另一种方法是,使用过渡的二值选择模型,这是因为在每一个离散时间区间中,两种结果都是可行的,即该时期要么结束,要么没有结束。

离散时间过渡模型的一般公式是:

$$\Pr[t_{a-1}\leqslant T<t_a\,|\,T\geqslant t_{a-1}\,|\,\mathbf{x}]=F(\lambda_a+\mathbf{x}'(t_{a-1})\boldsymbol{\beta}),\quad a=1,\cdots,A \tag{17.45}$$

这种设定是将回归元系数限制成随时间变化而为常值,对截距 λ_a 则限制成随时间变化而变动的,$a=1,\cdots,A$。函数 F 的一种明显选择是,标准正态 cdf 或逻辑斯蒂 cdf。于是,参数 λ_a 与 $\boldsymbol{\beta}$ 可通过叠放 logit 或叠放 probit 模型加以估计,其中,每一个持续期限都允许拥有各自的截距。这种方法由于简单而备受人们青睐。

所得到的似然函数是:

$$\mathrm{L}(\boldsymbol{\beta},\lambda_1,\cdots,\lambda_A)=\prod_{i=1}^{N}\Big[\prod_{s=1}^{a_i-1}(1-F(\lambda_s+\mathbf{x}_i'(t_{s-1})\boldsymbol{\beta}))\Big]\times F(\lambda_{a_i}+\mathbf{x}'(t_{a_i-1})\boldsymbol{\beta})$$

除对函数 F 选择以外,这类似于式(17.42),即离散时间 PH 模型的对数似然。风险(17.40)是在 $\ln\lambda_{0a}+\mathbf{x}(t_{a-1})'\boldsymbol{\beta}$ 处计算的极值 cdf,所以式(17.40)会产生互补双对数模型的二值选择模型(参见表 14.3),而不是更广泛使用的 logit 或 probit 模型。

17.11 持续期限失业例子

下面的实证例子运用了麦考尔(McCall,1996)的数据,布赖恩·麦考尔非常慷慨地向本书作者提供了他曾经研究的数据。这个数据集合来自 1986 年、1988 年、1990 年、1992 年的一月份当前人口调查的替代工人供给(DWS)。在这个例子中,我们把测量的持续期限(时期)称为失业持续期限,更准确地讲,它代表了无工作的持续期限,因为 DWS 并没有提供某个人是否寻找工作的信息。

就这种应用而言,需要关于取代后第一次工作是兼职的还是全日性情况的信息。为了确定取代后第一次工作是兼职的还是全日性的,采用下述方法。若某个实验者在调查时仍处于那份工作之中,同时若此实验者前一周在那份工作的每周工作小时数小于 35 个小时,则取代后的第一次工作被称为兼职的。

表 17.6 定义用于解释无工作持续期限的重要经济协变量。模型中估计的协变量数目相当大,但是这里仅列出重点关注的一个子集合。麦考尔(McCall, 1996)给出更为完整的描述。

表 17.6 失业持续期限:变量描述

变量名称	变量说明	均值
spell	无工作时期:2 周区间	6.248
CENSOR1	若以全日性工作再雇用,则为 1	0.321
CENSOR2	若以兼职工作再雇用,则为 1	0.102
CENSOR3	若再雇用却失去工作:工作状态未知	0.172
CENSOR4	若仍无工作,则为 1	0.375
UI	若提出 UI 申请,则为 1	0.553
RR	合格取代率	0.454
DR	合格忽视率	0.109
TENURE	占有年份丢失工作	4.114
LOGWAGE	周工资对数	5.693

失业持续期限以两周时间区间进行测量。引入四个二值变量(CENSOR1, CENSOR2, CENSOR3, CENSOR4)表示取代后第一次工作的状态。就本章的分析而言,我们使用 CENSOR1。因而,若某个人以全日性工作再雇用,则时期是完整的。另一个指示变量 UI 用作表示实验者是否提出失业申请。取代率,即丢失工作的每周救济金数量被每周工资数量去除,用变量 RR 来代表。"忽略"被定义成如下门限值,该门限值取决于得到兼职工作的失业保险的接收者在没有减少失业救济金条件下赚得的数值。忽略率是忽略被失去工作时的每周工资去除。在这个样本中,它是通过变量 DR 来描述的。正如我们所看到的,所有其他变量都是不言自明的。

我们以持续期限数据的描述分析开始。最简单的第一步是,画出卡普兰—迈耶生存曲线,如图 17.3 中的黑线所示。在估计卡普兰—迈耶生存曲线附近,细线代表 17.5.2 节曾研究过的 95%置信区间。正如人们所料,最初估计生存曲线迅速下降,然后缓慢地下降。

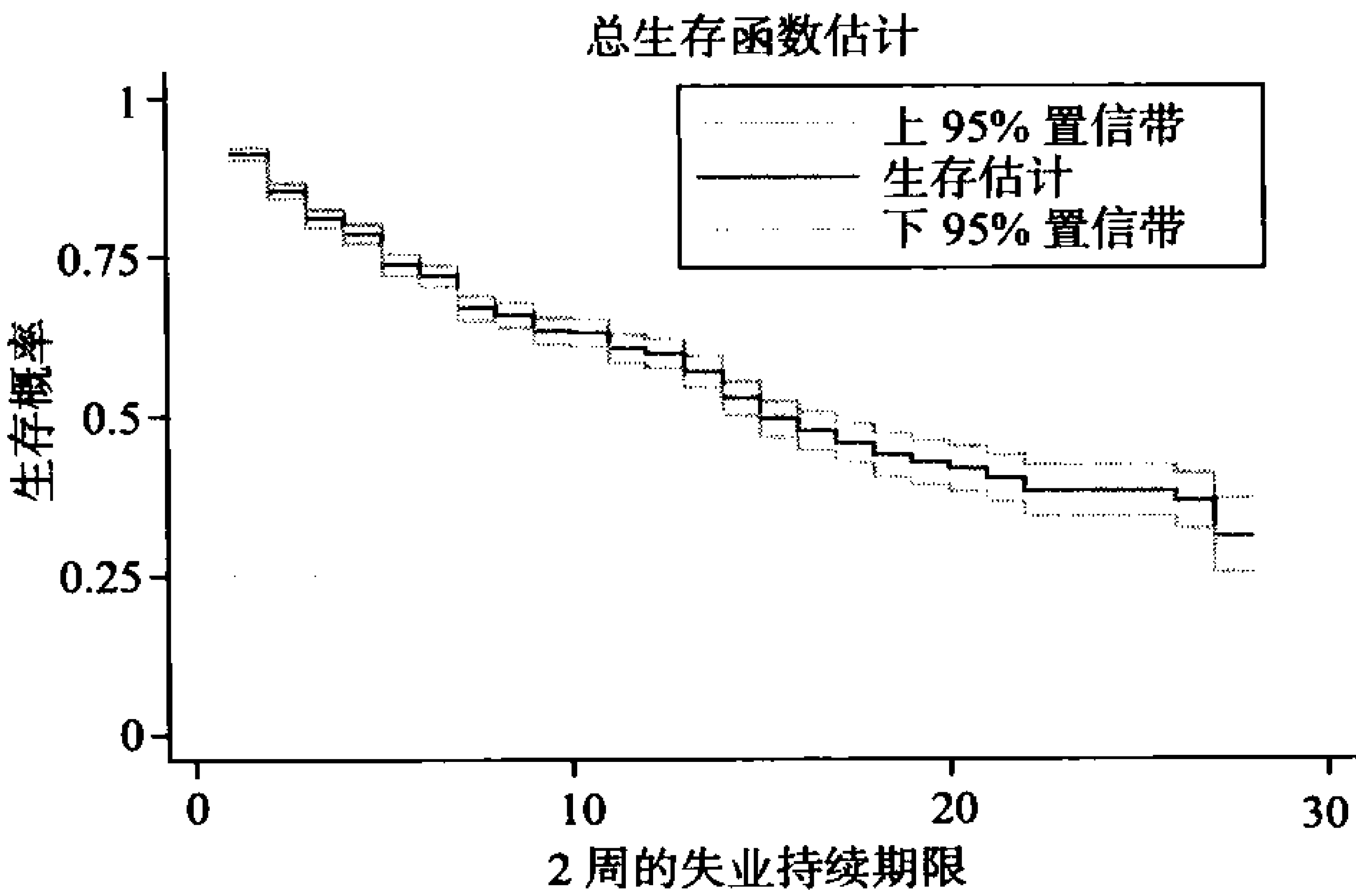


图 17.3 失业持续期限:借助于实验者是否接受失业保险而得出的估计生存函数。数据与图 17.3 中的一样。

考察表 17.7,可以发现,在第一个时期之后生存概率为 0.91,揭示被抽到的个体大致 9%在开始无工作时期的前两周结束了他们的期限。

表 17.7 失业持续期限

时间	生存函数	累计风险
1	0.912 1	0.087 9
2	0.854 1	0.151 4
3	0.810 3	0.202 7
4	0.786 4	0.232 2
5	0.737 6	0.294 3
⋮	⋮	⋮
12	0.597 4	0.500 5
13	0.568 0	0.549 6
14	0.527 0	0.621 9
⋮	⋮	⋮
26	0.365 1	0.980 9
27	0.309 8	1.132 5
28	0.309 8	1.132 5

在图 17.4 中,我们通过 UI 即实验者是否申请失业保险,画出生存函数。而且,正如人们所料,它表明与那些没有申请失业保险的人相比,申请失业保险的人更可能处于失业情况。

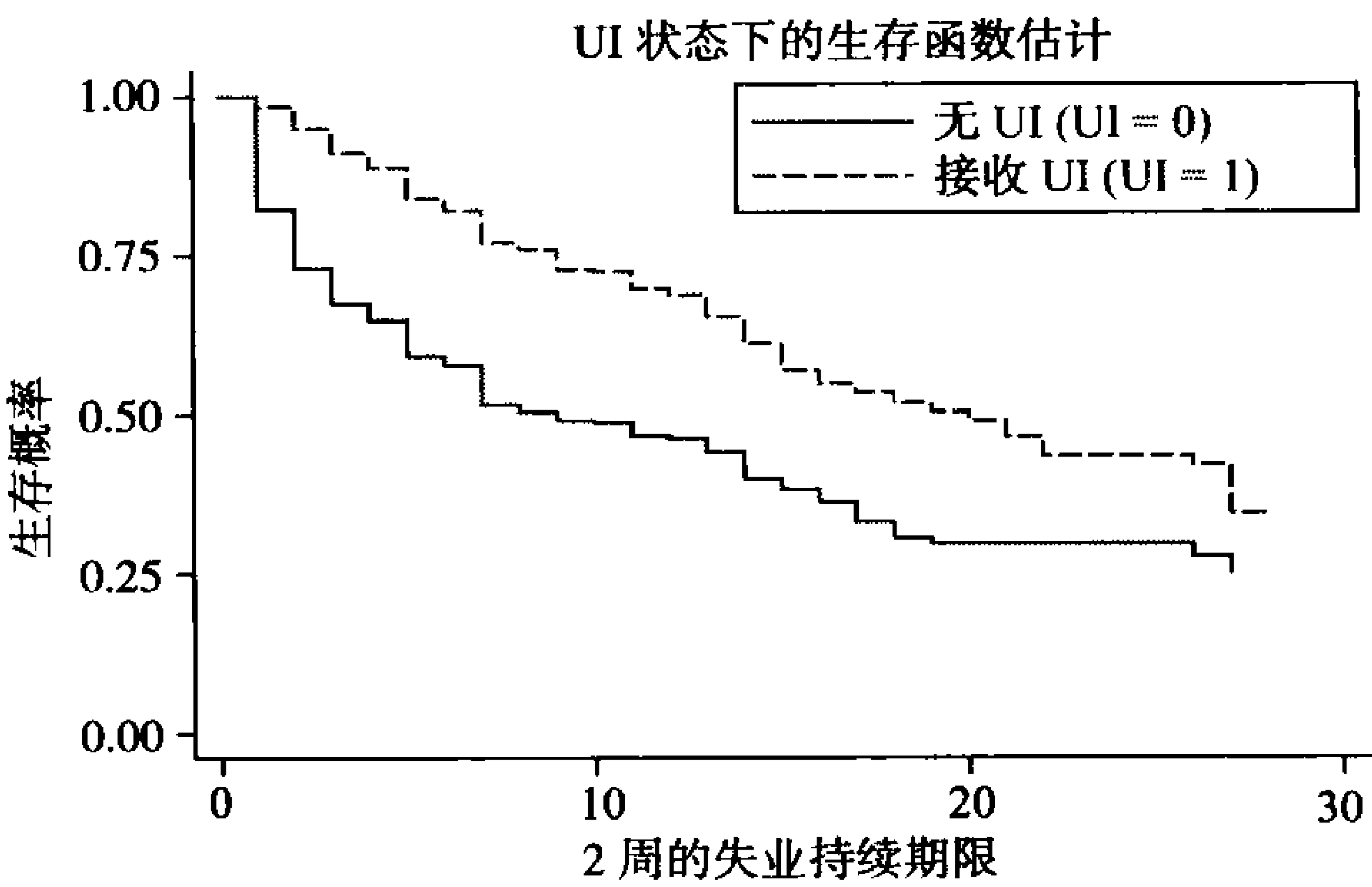


图 17.4 失业持续期限:借助于实验者是否接受失业保险而得出的估计生存函数。数据与图 17.3 中的一样。

图 17.5 中的纳尔逊—奥伦累积风险显示出风险率变异很小,可变换成近似线性风险。若未经整理的风险率变动很大,则累积风险表现了非线性特征。

通过 UI 接受所引起的累积风险函数揭示了预期模式,如图 17.6 所示:与那些申请失业保险的人相比,对未申请失业保险的人来说,此风险具有较高风险率。

下面,我们考察利用协变量 UI、RR、DR 和 LOG WAGE、交互作用项以及 34 个其他回归元的四种参数模型,表 17.8 与 17.9 没有报告出这 34 个其他回归元的系数。这四种类型是指数、威布尔、冈珀茨以及考克斯 PH 模型。将风险函数写成:

$$\lambda(t|\mathbf{x})=\lambda_0(t,\alpha)\phi(\mathbf{x},\boldsymbol{\beta})=\lambda_0(t,\alpha)\exp(\mathbf{x}'\boldsymbol{\beta})$$

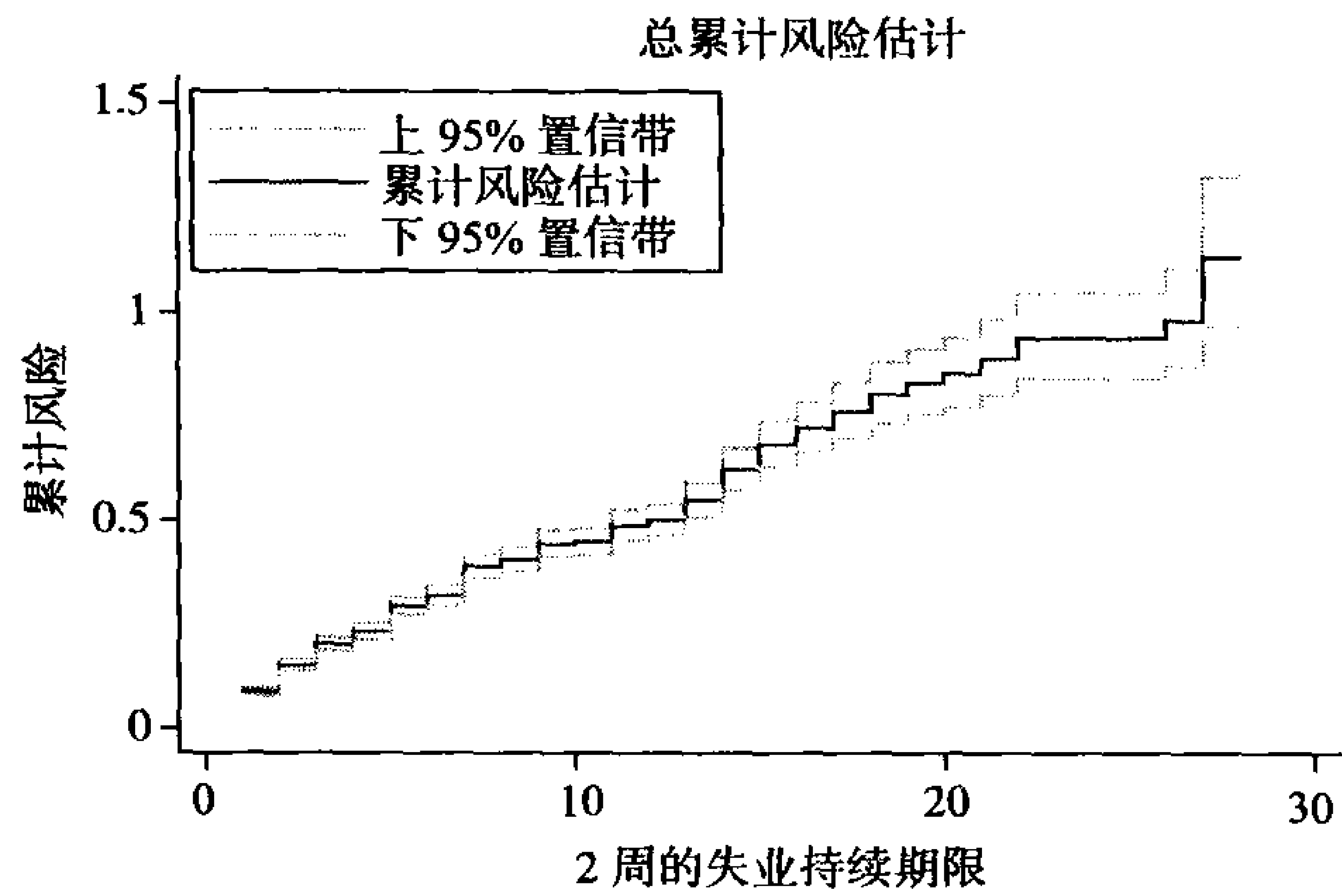


图 17.5 失业持续期限：累计风险函数估计的纳尔逊—奥伦估计。数据与图 17.3 中的一样。

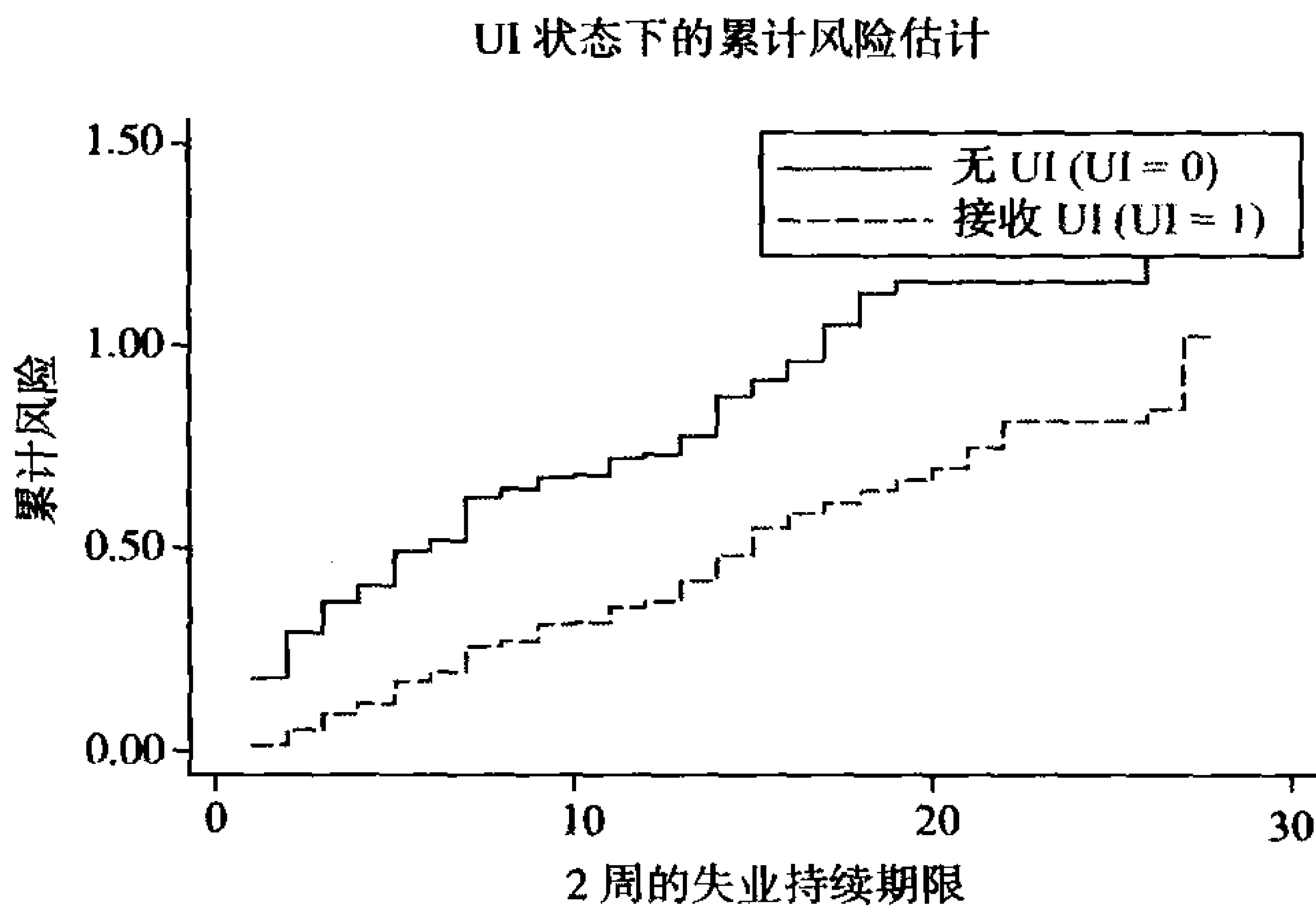


图 17.6 失业持续期限：借助于是否接受失业保险而得出的估计累计风险函数。数据与图 17.3 中的一样。

表 17.8 失业持续期限：由四种参数模型得出的估计参数

Var	指数		威布尔		冈珀茨		考克斯 PH	
	系数	t	系数	t	系数	t	系数	t
RR	0.472	0.79	0.448	0.70	0.472	0.78	0.522	0.91
DR	−0.576	−0.75	−0.427	−0.53	−0.563	−0.74	−0.753	−1.04
UI	−1.425	−5.71	−1.496	−5.67	−1.428	−5.69	−1.317	−5.55
RRUI	0.966	0.92	1.105	1.57	0.969	1.58	0.882	1.52
DRUI	−0.199	−0.20	−0.299	−0.28	−0.211	−0.21	−0.095	−0.10
LOGWANG	0.35	3.03	0.37	2.99	0.35	3.03	0.34	3.03
CONS	−4.079	−4.65	−4.358	−4.74	−4.097	−4.65	—	—
α			1.129					
−ln L	2 700.7		2 687.6		2 700.6		—	

表 17.9 失业持续期限:由四种参数模型得出的估计风险率

Var	指数		威布尔		冈珀茨		考克斯 PH	
	β	t	β	t	β	t	β	t
RR	1.603	0.63	1.565	0.57	1.604	0.62	1.686	0.71
DR	0.562	−1.02	0.653	−0.66	0.570	−0.99	0.471	−1.55
UI	0.241	−12.65	0.224	−13.12	0.240	−12.65	0.268	−11.53
RRUI	2.626	1.01	2.760	0.99	2.635	1.01	2.416	1.01
DRUI	0.819	−0.22	0.742	−0.33	0.810	−0.23	0.909	−0.10
LOGWANG	1.420	2.56	1.441	0.08	1.42	2.55	1.40	2.57
α			1.129					
$-\ln L$	2 700.7		2 687.6		2 700.6		—	

回顾,指数风险假定 $\lambda_0(t,\alpha)=$ 常值 $=\exp(a)$,对于某个常数 a,威布尔风险假定 $\lambda_0(t,\ \alpha)=\exp(a)\alpha t^{\alpha-1}$(即单调风险),冈珀茨风险假定 $\lambda_0(t,\alpha)=\exp(a)\exp(\gamma t)$,而考克斯 PH 模型没有截距,故对基准风险的形状没做什么假设。前面已经提及,这里的公式具有比例风险形式,并能被解释成参数回归模型或 AFT 模型。在这种似然函数的参数化中,参数$(\alpha,\boldsymbol{\beta})$是待估的。正如从表 17.8 看到的,威布尔模型的拟合显示正的状态相依性($\alpha=1.129>1$),也就是说,时期终止的概率随时期增长而增大。

对于考察的所有模型来说,只有 UI 与 LOGWAGE 是显著的,而其他协变量则是不显著的。就所有模型而言,UI 的估计系数都是负的,这蕴含那些申请失业保险的人无工作的时期终止得较慢。对不同模型来说,UI 的估计值变化很小;威尔布模型与冈珀茨模型中的 UI 估计值,在绝对值意义下,大致大于指数模型的 5%与 0.2%;而略小于考克斯 PH 模型的 8%。类似地,就所有模型而言,LOGWAGE的系数估计值都是正值,这显示该值在各种不同模型上变化非常小。

在生物统计学中,各种不同参数化经常基于 PH 模型来使用,而在经济计量文献中,普遍做法是报告 AFT 模型的风险函数系数$(\alpha,\boldsymbol{\beta})$的估计值。注意到,风险比率 $\lambda(t|\mathbf{x})/\lambda_0(t,\alpha)=\phi(\mathbf{x},\boldsymbol{\beta})=\exp(\mathbf{x}'\boldsymbol{\beta})$。对于类别 0/1 的标量变量 x,从 0 到 1 的变动影响由 $\exp(\beta)-1$ 给出,这测量了相对于基准风险的影响。许多软件都为用户提供选择,要么用上述两种变量之一估计模型,要么同时用这两种度量估计模型的选择。两种参数化的相对优点,已由克利夫斯、古尔德和吉特莱斯(Cleves, Gould and Guitirrez, 2002)讨论过。

考察表 17.9 中的指数设定,其系数是对应于表 17.8 中的指数。这里,UI 具有风险比率 0.241。这意味着,属于申报失业保险的实验者类型减少了基准风险的大致 76%风险。类似地,对于威布尔、冈珀茨以及考克斯 PH 模型,风险分别减少了大致 78%、76%以及 73%。

对于这个例子,我们考虑右删失,并忽略不可观测异质性的作用。因此,从三种模型得出的结果,在性质上相似。不过,被包括进来的相对极少的几个变量系数都是显著的,这一点或许表明,没有被解释的大变异(也许由不可观测异质性引起)

是一个严重问题。此问题由下一章进一步研究。

17.12 应用研究

绝大多数计算机软件包都提供参数生存分析计算机程序的良好选择。可广泛利用标准的非参数卡普兰—迈耶生存函数估计，既有置信区间的，也有无置信区间的；既有数值输出的，又有图形输出的。在一些情况下，生存分析模块依据特定手册表现得充分详细。例如，阿莉森(Allison, 1995)提出 SAS 系统的生存分析实用指南；克利夫斯(Cleves, 2002)等人给出 STATA 的指导式生存分析指南。这些指南不仅解释执行特殊程序命令的原理，而且在许多情况下，它们提出源于特定数据特性、可供选择参数化以及对结果诠释的副标题的宝贵见解。学习持续期限数据分析的一种方便途径是，通过利用经济计量学或统计软件包诸如 LIMDEP、STATA、SAS 或 S-Plus 里面的例子加以学习。程序手册本身也是一种标准模型的良好信息来源。

17.13 文献注释

17.3～17.7 卡尔布弗莱舍和普伦蒂斯(Kalbfleisch and Prentice, 1980, 2002)的文献已是生存分析方面的经典统计文献，他们尤其强调考克斯模型。其他一些有用的文献包括劳利斯(Lawless, 1982)、考克斯和奥克斯(Cox and Oakes, 1984)，以及现在出版的大量生存分析的统计学教科书。关于贝叶斯的研究，参见易卜拉欣、陈和森哈(Ibrahim, Chen, and Sinha, 2001)。最近，强调计数过程方法的统计研究日益增多，详细内容参见弗莱明和哈林顿(Fleming and Harrington, 1991)、安德森等人(Andersen et al., 1993)。

这些文献都非常具有挑战性，尤其是后者。兰开斯特(Lancaster, 1990)提供生存分析的一个详尽研究，尽管表述形式却是相当具有技术性的，而且该书更倾向于后两章所述的过渡内容的一般性专题。对于社会科学家来说，像兰开斯特一样，阿利森(Allison, 1984)的优秀解释涵盖了多于单时期的生存分析。对微观经济计量学实践者来说，由基弗(Kiefer, 1988)撰写的综述则是一个良好的开端。

17.8 对于偏似然法，兰开斯特(Lancaster, 1990)已经给出一个深入透彻的讨论。

17.10 关于离散风险函数，迈耶(Meyer, 1990)、哈恩和豪斯曼(Han and Hausman, 1990)以及布莱克等人(Blake et al., 1990)的文献都是有益的。这些文章一般地考虑了不可观测异质性，下面一章将讨论这个专题。

17.11 基弗(Kiefer, 1988)与格林(Greene, 2003)曾列举一些经济应用。参数简化式形式持续期间分析的优秀例子是，由兰开斯特(Lancaster, 1979)、纳伦德拉内森、尼克尔和斯特恩(Narendranathan, Nickell and Stem, 1985)、贾吉娅(Jaggia, 1991c)以及格里茨(Gritz, 1993)给出。目前，研究重点转向计算结构更加复杂的持续期限模型。范登堡(Van den Berg, 1990)与费拉尔(Ferall, 1997)已经给出

一些例子。持续期限分析的绝大多数应用都是简化式模型。经济学家提出结构持续期限模型：参考文献包括兰开斯特（Lancaster，1990）与范登堡（Van den Berg，2001）。范登堡还提供了 PH 模型经济理论基础的有趣讨论。持续期限数据经常利用各种等待时间的概念加以分析。图纳勒和普里切特（Tunali and Pritchett，1997）曾经运用三种可选择的概念：日历时间、年代以及持续期限。

习 题

17－1 ［改编自萨普兰（Sapra，1998）。］证明，第一类帕累托密度为 $f(t)=\alpha k^{\alpha}/t^{\alpha+1}$，$\alpha>0$，$t\geqslant k\geqslant 0$ 的持续期限数据模型是一种加速失效时间持续期限模型，但它不是比例风险模型。［提示：证明，$\ln t$ 可被表示成关于 $k=\exp(\mathbf{x}'\boldsymbol{\beta})$ 的线性函数，具有可加异方差误差。］

17－2 ［依据兰开斯特（Lancaster，1979）。］对于下述每一种情况，利用持续期限密度 $f(t|\mathbf{x},\boldsymbol{\theta})$ 与生存函数 $S(t|\mathbf{x},\boldsymbol{\theta})$，研究 N 个观测值的联合似然的适当表达式。

（a）可以利用独立的完整持续期限 t_i 的样本，$i=1,\cdots,N$。

（b）样本由下述方式生成。最初，一些个体是失业者与被采访者的混合体。然后，他们在 h 个时期之后被重新采访。选出的个体失业了 t 个周。在选择与采访之间，有些人找到了工作，而另一些人则没找到工作。对于找到工作的那些人来说，失业时期的终止时间是已知的。

（c）此情况与（b）情形一样，只是人们并不知道失业时期何时终止。

17－3 （a）利用麦考尔数据集的 50％随机样本，通过删失类型，即考虑过渡到全职还是兼职的就业形式，估计卡普兰—迈耶非参数生存与综合风险函数。

（b）若忽略时期终止形式的删失变量，在下面参数分布假设下：（i）指数；（ii）威布尔；（iii）对数 logistic；（iv）考克斯 PH，估计失业持续期限的风险模型。这里所用的协变量与本章中的一样。

（c）比较模型（i）～（iii），并讨论哪一个会提供对数据的最佳拟合。就失业时期的持续期限独立性（风险函数的形状）而言，每一个模型都蕴含什么内容？

18

混合模型与不可观测异质性

18.1 引 论

存在大量统计文献与经济计量文献，涉及不可观测异质性这个专题。观测异质性意指回归元所测量的个体间的差异，而不可观测异质性意指所有其他的差异。这两种因素都会影响到生存时间。在存在不可观测异质性的条件下，甚至具有全部协变量的相同值的个体，在离开已知状态时也可能具有不同的风险。当人们忽略不可观测异质性时，它的影响就会与基准风险的影响相混淆。

为了进一步研究，考虑一个著名的实证例子。人们已经知道，来自失业的总风险率是一个关于失业时期长度的下降函数。倘若所有个体均是相同的，则这蕴含负的持续期间相依性，即离开失业的下降概率会使个体继续失业的时间较长。不过，假定失业总体中有两种不同类型的个体，一种是 F 类型(快的)，它具有常值风险率 0.4，而另一种是 S 类型，它的常值风险率为 0.1。总体由两种类型 50/50 混合而成。那么，对于 100 个 F 类型人员，我们观察到在第一个时期有 40 个过渡，第二个时期有 24 个过渡，而第三个时期有 14.4 个过渡。对于 S 类型，我们在第一个时期、第二个时期和第三个时期分别观测到 10、9 和 8.1 个过渡。因此，总的过渡比例分别是 $(40+10)/200=0.25$、$(24+9)/150=0.22$ 和 $(14.4+8.1)/117=0.192$，这表明，下降总风险是各个异质性组加总的结果，各组自身为常值，却具有不同的风险率。准确表述持续期限独立性，则需要并入不可观测异质性的模型。

在线性回归模型中，假如异质性与回归元是独立的，则不可观测异质性不会引起复杂问题。在这种情况下，条件均值没有变动，不可观测异质性被并入误差项之中，从而不存在省略变量偏倚。与之相比，不可观测异质性在持续期模型中通常会引起一些问题。在最简单模型中，诸如指数模型，可能要设定与回归元不相关的乘法不可观测异质性，以此使条件均值持续限期没有变化。不过，甚至在最简单情况下，条件风险函数确实变化了，而当已知存在删失，并且例如已知政策制定者的关注内容在于，确定退出失业率如何随失业时期长度而变动，而且条件风险函数从必要性考虑要加以建模。

不可观测异质性的作用，在大量实证研究中占据着探索令人困惑之谜与问题的核心境地。尽管本章关注于持续期限模型，但大部分问题仍具有更一般的意义。

而且，这里所考虑的内容及方法与所有的经济计量模型有关，因为所有经济计量模型都会从模型中省略某种特定个体的无法观测变量。其他章节的一些重要例子，包括随机函数 logit(15.7 节)、样本选择(16.4 节)、计数的有限混合(20.4 节)以及面板数据的固定效应与随机效应(第 21～23 章)。这些因素被归入不可观测异质性专题之中。在生物科学中，还使用脆弱性(**frailty**)术语。在实际研究中，(乘法)不可观测异质性测量风险率(死亡率压力)的增大或减小，对已知个体相对于平均水平个体而言产生影响。特定个体异质性不必是时不变的，但在横截面模型里，对它这样假设就很适宜。

重要的是，考察这类不可避免错误设定的后果。由普通线性多元回归分析知道，一般地讲，此类省略可以产生省略偏倚。在持续期限模型中，作为非线性和不可观测异质性的分析就显得更加复杂。引入不可观测异质性导致了所谓混合模型(**mixture models**)中的一类重要形式，混合模型只是此类众多模型称谓之一。本章的论题，既涉及对混合模型的生成与分析，又讨论省略异质性所引致的严重后果。

对异质性与真实状态相依性进行辨别已是一个悠久问题，对它可追溯到对关于真实与表面传染加以探讨的历史。内曼(Neyman)确信他早期的观点：纵向数据可能从经验上讲对实施这种辨别是必需的。不过，仅仅利用横截面数据时，会有严重依赖于强参数假设的倾向。最新的研究文献突出了使经验分析无此类假设之忧，并对支持模型假设的有效性进行探索。

本章第一部分是 18.2～18.4 节，研究基于异质性连续分布的混合模型。18.5 节阐述基于离散异质性的模型。18.6 节考虑来自流动数据与存量数据的两种不同持续期限概念之间的关系。错误设定的检验以及忽略异质性问题，将在 18.7 节讨论。18.8 节的实证例子阐明了本章探索的几种思想。

18.2 不可观测异质性与离散度

本节关注指数模型与威布尔模型中的不可观测异质性。我们考察如下乘法形式的不可观测异质性，通过积分去掉它后，使条件均值没有变化，但又不会使条件方差变大，更重要地讲，并没有导致条件风险函数的变化。同样，对盛行的服从伽玛分布异质性的威布尔模型加以阐述。

18.2.1 *混合*

要考察的最简单模型是指数持续期限模型。在不含异质性的指数回归中，完整时期 t_i 分布被设定成以可观测的弱外生协变量 $\mathbf{x}_i$ 为条件。这等价于对条件均值函数设定成非随机的：$\mathrm{E}[T|\mathbf{x}]=\exp(\mathbf{x}'\boldsymbol{\beta})$。在混合模型里，我们反而对 $(t_i|\mathbf{x}_i, \nu_i)$ 分布加以设定，其中，附加的 ν_i 表示第 i 个观测值的不可观测异质性。简单地讲，个体被假定成随机地不同于以不能完全由观测协变量加以解释的方式。t_i 的边缘分布，可通过对 ν_i 进行平均来获得。

必须对联系 t_i 与 $(\mathbf{x}_i, \nu_i)$ 的准确函数形式加以设定。一种广泛运用的函数形式是，具有乘法误差的指数均值。例如，考察具有不可观测异质性的 PH 模型。由

17.8 节，我们具有比例风险模型(17.25)与(17.26)，为了包含乘法项 ν 而对两个模型加以推广。也就是说：

$$\lambda(t|\mathbf{x},\nu)=\lambda_0(t)\exp(\mathbf{x}'\boldsymbol{\beta})\nu,\ \nu>0$$

因此，我们能获得如下综合基准风险的表达式：

$$\lambda_0(t)=\lambda(t|\mathbf{x},\nu)\exp(-\mathbf{x}'\boldsymbol{\beta})\nu^{-1} \tag{18.1}$$

$$\int\lambda_0(u)du=\exp(-\mathbf{x}'\boldsymbol{\beta})\nu^{-1}\int\lambda(u|\mathbf{x},\nu)du$$

$$\ln\left[\int\lambda_0(u)du\right]=-\mathbf{x}'\boldsymbol{\beta}-\ln\nu+\varepsilon$$

其中，$\varepsilon=\ln\int\lambda(u|\mathbf{x},\nu)du$ 被假定成与回归元独立的且具有删失时间形式的。一个普遍正规化约束是 $\mathrm{E}[\nu]=1$。当 $\nu>1$ 时，风险率大于平均实验者；当 $\nu<1$ 时，风险率小于平均实验者。独立性假设显得太强，而且未必现实。同样地，乘法异质性假设也是特设的，不过与可加误差相比，这种形式在数学上处理方便，备受人们青睐，可能违背 t_i 的非负性。一种标准方法涉及对 ν_i 分布进行假定，然后推导出 t_i 边缘分布。

乘法异质性具有两个重要且有关联的结果。并不令人感到惊讶的是，混合分布的(以可观测变量为条件的)方差大于其母分布的方差(以可观测变量和异质性为条件的)。也就是说，使方差变大。考察指数均值情况。用：

$$\begin{aligned}\mu_i^*&=\mathrm{E}[t_i|\mathbf{x}_i,\nu_i]\\&=\exp(\mathbf{x}_i'\boldsymbol{\beta})\nu_i\\&=\exp(\mathbf{x}_i'\boldsymbol{\beta})\exp(\varepsilon_i)\\&=\exp(\beta_0+\varepsilon_i+\mathbf{x}_{1i}'\boldsymbol{\beta}_1)\end{aligned} \tag{18.2}$$

代替 $\mu_i=\exp(\mathbf{x}_i'\boldsymbol{\beta})$，其中，第三行的不可观测异质性项 ν_i 被重新定义成 $\exp(\varepsilon_i)$，而最后一行里的 $\mathbf{x}_i'\boldsymbol{\beta}$ 项被分成截距项与斜率项。最后一行可解释成条件均值具有一种随机变化的截距($\beta_0+\varepsilon_i$)。通常假定，ν_i 是 iid 的，可能服从已知函数的分布，同时 ν_i 与 $\mathbf{x}_i$ 是独立的。

假定 ν_i 是 iid，满足 $\mathrm{E}[\nu_i]=1$ 与 $\mathrm{V}[\nu_i]=\sigma_\nu^2$。这里假设 $\mathrm{E}[\nu_i]=1$ 使得对截距识别成为可能。对指数密度来说，可将 t_i 的矩推导成 $\mathrm{E}[t_i|\mathbf{x}_i,\nu_i]=\mu_i\nu_i$，并利用 A.8 节关于方差分解的结果，得出：

$$\begin{aligned}\mathrm{V}[t_i|\mathbf{x}_i]&=\mathrm{V}_\nu[\mathrm{E}_{t|\nu,\mathbf{x}}(t_i|\nu_i,\mathbf{x}_i)]+\mathrm{E}_\nu[\mathrm{V}_{t|\nu,\mathbf{x}}(t_i|\nu_i,\mathbf{x}_i)]\\&=\mu_i^2\mathrm{V}(\nu_i)+\mu_i^2(\mathrm{V}(\nu_i)+1)\\&=\mu_i^2[1+2\sigma_\nu^2]\\&>\mu_i^2\end{aligned} \tag{18.3}$$

不可观测异质性促使无条件方差变大。

18.2.2 *选择异质性分布*

考察 t 的分布如何受异质性影响。这要求我们从 $S(t|\mathbf{x},\nu)$ 通过积分去掉异质

性项 ν 来考虑 t_i 的分布。通常，ν 的参数分布是人们设定的。选择这种分布时需注意什么呢？

为了保持性质 $\nu_i>0$，我们用正直线的支集来设定分布。一些例子包括，伽玛、逆高斯以及对数正态。

伽玛密度(**gamma density**)是：

$$g(\nu;\delta,k)=\frac{\delta^k\nu^{k-1}\exp(-\delta\nu)}{\Gamma(k)},\quad \nu>0 \tag{18.4}$$

它具有 $E[\nu]=k/\delta$ 且 $V[\nu]=k/\delta^2$。若进行正规化，则令 $k=\delta$，从而 $E[\nu]=1$，$V[\nu]=1/\delta$。从数学形式上讲，伽玛假设方便。就持续期限建模而言，它也用于一系列流行软件包之中。

逆高斯密度(**inver-Gaussian density**)是：

$$g(\nu;\delta,\theta)=\delta\pi^{-1/2}\exp(2\delta\theta^{1/2})\nu^{-3/2}\exp(-\theta\nu-\delta^2/\nu),\quad \nu>0 \tag{18.5}$$

它具有 $E[\nu]=\delta\theta^{-1/2}$ 且 $V[\nu]=\delta\theta^{-3/2}/2$。若进行正规化 $\theta=\delta^2$，则得到 $E[\nu]=1$ 且 $V[\nu]=1/2\theta$。相对于伽玛情况来说，逆高斯分布有较大的尾部概率。

这些不一定产生解析形式上容易处理的 t 的边缘分布。正如将要看到的，某些组合诸如指数与伽玛，或者威布尔与伽玛都会得到闭形式边缘分布，而另一些组合则不会得出。不过，这种考虑仅仅是数学与计算机上的方便而已，因此不必非得这样做。不幸的是，人们很少有源自经济理论的持续期限建模方面的指南。

第二个考虑是一般性与灵活性。伽玛模型具有相当灵活性，并有许多引人注目的性质。然而，逆高斯模型可能更好地处理宽尾分布。这两种模型都是单参数族(正规化之后)。霍高(Hougaard，1986)引进了更为灵活的两个参数族，该族具有伽玛与逆高斯作为特殊情况的性质。本章稍后还提供相当灵活的离散(非参数)表示式。

18.2.3 *威布尔—伽玛混合*

其次，我们考虑流行的威布尔—伽玛混合(**Weibull - gamma mixture**)，它能被特定化为指数—伽玛情况。这种模型是混合比例风险(MPH)模型的一个重要特例。当然，威布尔—伽玛混合具有独立的关注内容，这是因为它拥有较大的灵活性，尤其是可以证明，它既包含递增风险又包含递减风险。

关于威布尔模型，以乘法形式 ν 为条件的生存函数是：

$$S(t|\nu)=\exp(-\mu t^{\alpha}\nu),\quad \lambda>0,\alpha>0 \tag{18.6}$$

其中，μ 代替第 17 章使用过的 α。

无条件生存函数是由平均生存函数给出的。利用 ν 的密度 $g(\nu)$ 作为加权函数，对异质性总体加以平均，得出：

$$S(t)=E_{\nu}[S(t|\nu)]=\int S(t|\nu)g(\nu)d\nu \tag{18.7}$$

对 $g(\nu)$ 的不同选择产生了各种不同的混合形式。在既有连续分布又有离散分布

的情况下，只需对解释做出适当变化就有效。式(18.7)的积分可能没有解析解。例如，若 $g(\nu)$ 是对数正态密度，则积分确实没有解析解，但如果它是伽玛分布，就存在解析解。为了数学处理方便，我们以下述的伽玛情况开始研究。

已知伽玛异质性，无条件生存函数是：

$$S(t)=\int_0^{\infty}\exp(-\mu t^{\alpha}\nu)\frac{\delta^k\nu^{k-1}\exp(-\delta\nu)}{\Gamma(k)}d\nu \tag{18.8}$$
$$=\frac{\delta^k}{\Gamma(k)}\int_0^{\infty}\nu^{k-1}\exp(-\nu(\mu t^{\alpha}+\delta))d\nu$$

为了获得混合密度，我们要求解此积分。设 $\mu t^{\alpha}+\delta=\beta$，得出：

$$S(t)=\frac{\delta^k}{\Gamma(k)}\int_0^{\infty}\frac{(\nu\beta)^{k-1}}{\beta^{k-1}}\exp(-\nu\beta)d\nu$$

定义 $y=\nu\beta$，因而 $d\nu=\beta^{-1}dy$，并且：

$$S(t)=\frac{\delta^k}{\Gamma(k)\beta^k}\int_0^{\infty}y^{k-1}\exp(-y)dy$$
$$=\frac{\delta^k}{\Gamma(k)}\frac{\Gamma(k)}{(\mu t^{\alpha}+\delta)^k}$$
$$=\delta^k(\mu t^{\alpha}+\delta)^{-k}$$
$$=[1+(\mu t^{\alpha}/\delta)]^{-k} \tag{18.9}$$

其中，第二行用到 $\Gamma(k)$ 的定义，并代入 β 而得到。

无条件持续期限密度可通过对 t 求微分，并用 -1 乘获得，从而：

$$f(t)=\frac{k}{\delta}\mu\alpha t^{\alpha-1}[1+(\mu t^{\alpha}/\delta)]^{-(k+1)} \tag{18.10}$$

无条件风险函数 $\lambda(t)=f(t)/S(t)$ 是：

$$\lambda(t)=\frac{k}{\delta}\mu\alpha t^{\alpha-1}[1+(\mu t^{\alpha}/\delta)]^{-1} \tag{18.11}$$

通过设置 ν 的均值为 1，对这些一般表达式加以特定化研究，也就是说，设 $k=\delta$，正规化为 $\mathrm{E}[\nu]=1$，从而得到下述威布尔—伽玛混合形式的表达式：

$$S(t)=[1+(\mu t^{\alpha}/\delta)]^{-\delta} \tag{18.12}$$

$$f(t)=-\frac{\partial S(t)}{\partial t}=\mu\alpha t^{\alpha-1}[1+(\mu t^{\alpha}/\delta)]^{-(\delta+1)} \tag{18.13}$$

$$\lambda(t)=-\frac{\partial\ln S(t)}{\partial t}=\mu\alpha t^{\alpha-1}[1+(\mu t^{\alpha}/\delta)]^{-1} \tag{18.14}$$

当方差 $1/\delta$ 趋于 0 时，就趋于威布尔风险函数。

威布尔模型允许出现递增风险或者递减风险，只是从约束形式上看，要假定在个体水平上拥有条件单调风险。不过，这种混合分布在经济计量学文献中仍然十分盛行，主要因为它具有方便的性质，参见兰开斯特(Lancaster, 1979)，纳伦德拉内森、尼克尔和斯特恩(Navendranathan, Niclcell, and Stern, 1985)。

为了专门研究指数—伽玛混合的结果，设 $\alpha=1$。从而得出 $S(t)=[1+(\mu t/\delta)]^{-\delta}$，$f(t)=\mu[1+(\mu t/\delta)]^{-(\delta+1)}$，而且 $\lambda(t)=\mu[1+(\mu t/\delta)]^{-1}$。指数—伽玛混合分布，即著名的第二类帕累托分布，与指数分布相比，其尾部拥有更大的质量。两者之间的区别依赖于方差 $1/\delta$。只有当 $\delta>r$ 时，r 阶矩才存在。

18.2.4 *混合风险函数的解释*

经济应用中的一个重要问题是，持续期限相依性在持续期限数据中是正的还是负的。例如，当失业时期长度增加时，退出失业的概率是会变大（比如，因为工人的保留工资下降），还是会变小（比如，工人被看成是有害商品）呢？在 iid 情况下，这很容易利用非参数估计方法建立起来。不过，对于非 iid 情况，原始数据中的递减风险归因于对各个不同个体进行加总，其中，每一个个体都具有独立的常值风险值，或者归因于每一个个体的递减风险。对这两者情况加以区分很困难。

在指数伽玛混合条件下，考察存在不可观测异质性时对风险函数进行解释的问题。注意到，甚至如果个体风险（比如，以 ν 为条件的风险）在 μ 处为常值，那么对风险 $\lambda(t)$ 的平均或加总关于 t 是向下倾斜的。这并不意味着，个体风险率存在负的持续期限相依性。更准确地讲，这是通过对那些风险率随机地存在差别的个体进行加总而引起的。类似的不正确解释也出现在威布尔伽玛情况中。在这种情况下，风险函数的真实斜率依赖于 α，但平均或总风险函数的斜率却受到异质性存在的影响。因而，对不可观测异质性的忽略，会导致对风险函数斜率的低估。这个结果看起来相当一般[参见兰开斯特(Lancaster, 1990)]。萨伦特(Salant, 1977)给出了对这种现象的早期推广讨论。

这个结果是下面陈述的基础[比如，参见兰开斯特(Lancaster, 1979)；赫克曼和辛格(Heckman and Singer, 1984a)]：在忽略不可观测异质性条件下，对风险函数的估计可能产生严重偏倚。我们的讨论完全出于对风险模型中不可观测异质性的检验。在威布尔混合模型背景下，考虑 $S(t)=\int \exp(\mu t^{\alpha}\nu)g(\nu)d\nu$ 的自变量。总风险函数是：

$$\begin{aligned}\lambda(t)&=-\int\frac{\partial \ln S(t|\nu)}{\partial t}g(\nu)d\nu\\&=\alpha\mu t^{\alpha-1}\int\frac{\nu\exp(-\mu t^{\alpha}v)}{S(t|\nu)}g(\nu)d\nu\\&=\alpha\mu t^{\alpha-1}\mathrm{E}[\nu|T\geqslant t]\end{aligned}$$

因为 $\mathrm{E}[\nu|T\geqslant t]$ 是针对那些在时间 t 时生存 ν 的平均值，所以当有较大 ν 值的个体比有较小 ν 值的个体更快地离开此状态时，它一定随时间而递减。这引起了总风险函数的斜率变化。此种现象也被认为是选择性偏倚(**selectivity bias**)（16.5 节）的形式。正式地讲，ν 关于时间的均值可写成：

$$\mathrm{E}[\nu|T\geqslant t]=\int\frac{\nu\exp(-\mu t^{\alpha}\nu)}{S(t|\nu)}g(\nu)d\nu$$

因此，就威布尔混合模型而言：

$$\frac{\partial \mathrm{E}[\nu | T \geqslant t]}{\partial t} = -\alpha\mu t^{\alpha-1}\left[\int \frac{\nu^2 \exp(-\mu t^\alpha \nu)}{S(t|\nu)} g(\nu) d\nu\right] + \alpha\mu t^{\alpha-1}\left[\int \frac{\nu \exp(-\mu t^\alpha \nu)}{S(t|\nu)} g(\nu) d\nu\right]^2$$
$$= \alpha\mu t^{\alpha-1}\{\mathrm{E}[\nu^2 | T \geqslant t] - (\mathrm{E}[\nu | T \geqslant t])^2\}$$
$$= \alpha\mu t^{\alpha-1}\mathrm{V}[\nu | T \geqslant t] \tag{18.15}$$
$$< 0$$

所以，忽略异质性导致估计风险率比真实风险率下降得更快或上升得更慢。

在没有异质性与有异质性的模型之间进行另一个有意思的比较是，考察协变量变化对风险率的比例影响。在不存在异质性条件下，有：

$$\ln \lambda(t|\mu) = \ln \lambda(\mu t^{\alpha-1}) + \ln \alpha$$

而 x_j 的变化对 μ 的比例影响是：

$$\frac{\partial \ln \lambda(t|\mu)}{\partial x_j} = \beta_j$$

这是比例风险模型的性质。

考虑到不可观测异质性时，有：

$$\ln \lambda(t|\mu) = \ln \lambda(\mu t^{\alpha-1}) + \ln \alpha + \ln \mathrm{E}[\nu | T \geqslant t]$$
$$= \ln \alpha + \ln \mu + (\alpha-1)\ln t + \ln \mathrm{E}[\nu | T \geqslant t]$$

注意，$\ln \mu = \mathbf{x}'\boldsymbol{\beta}$ 与 $\partial \mathrm{E}[\nu | T \geqslant t]/\partial x_j = -\mu t^\alpha \mathrm{V}[\nu | T \geqslant t]\beta_j$，就威布尔混合模型而言可得：

$$\frac{\partial \ln \lambda(t|\mu,\nu)}{\partial x_j} = \beta_j\left[1 - \frac{\mu t^\alpha \mathrm{V}[\nu | T \geqslant t]}{\mathrm{E}[\nu | T \geqslant t]}\right] \tag{18.16}$$
$$< \beta_j$$

此结果表明，已知异质性，x_j 变化的比例影响较小但依赖于 t，因而不再是比例风险形式。所以，由模型得出的估计可能会导致错误，甚至当不可观测异质性项与所包括的协变量并不相关时。

对于比威布尔模型更一般的模型来说，兰开斯特和尼克尔（Lancaster and Nickell，1980）曾经讨论了不可观测异质性的类似结果。

18.3 混合模型的识别

与混合模型有关的问题是一般的识别问题（**identification problem**）。已知单时期的观测数据（t，$\mathbf{x}$），这个问题涉及把个体贡献分解成基准风险的平均生存概率、不可观测异质性以及协变量的逻辑可能性。更明确地讲，当 PH 模型是不可识别时，将个体贡献分解成持续期限相依性与不可观测异质性在逻辑上行不通。如同大部分识别讨论一样，要对公式施加某些约束。在经济计量文献中，已经对（混合）比例风险情况进行了详细研究。赫克曼和辛格（Heckman and Singer，1984b）、

埃尔贝斯和里德(Elbers and Ridder, 1982)在某些条件下建立起 MPH 模型的识别。范登堡(Van den Berg, 1982)提供了这些早期证明及后来探索研究贡献的优秀讨论。

对 MPH 模型识别的讨论开始于平均或加总生存函数(**average or aggregate survivor function**):

$$S(t|\mathbf{x}) = \mathrm{E}_\nu[S(t|\mathbf{x},\nu)] \qquad (18.17)$$
$$= \int \exp(-\nu\Lambda_0(t)\phi(\mathbf{x}))g(\nu)d\nu$$

这假定了如同式(18.1)的比例风险,使用17.8节的PH公式,却没有对 Λ_0、ϕ 或者 g 做出假设。此处,$\Lambda_0(t) = \int_0^T \lambda_0(s)ds$。当已知数据时,如果函数 λ_0、g 和 ϕ 都是唯一的,那么称该模型在非参数形式上是可识别的。我们加上修饰语"在非参数形式上",正是因为对函数形式没有做任何假设。

观测到的生存时间出现变异,这归因于协变量 $\mathbf{x}$、ν 以及持续期限相依性函数(基准风险)上的变化。识别性意味着该变化有唯一分解。对识别性进行证明必须表明,这些各自分解在原则上是可识别的。大部分可利用的证明,都运用高等数学工具去验证似然函数可唯一地被分解。梅利诺和末吉(Melino and Saeyoshi, 1990)已经给出一个较简明的证明。

非参数估计所要求的条件如下:(i) 异质性项 ν 被假定成时不变的,且与 $\mathbf{x}$ 分布是独立的;(ii) $g(\nu)$是非退化的,并具有有限均值(即 $\mathrm{E}[\nu]<\infty$);(iii) $\phi(\mathbf{x})>0$,对于所有 $\mathbf{x}$;(iv) $\lambda_0(t)$在$[0, \infty)$上连续且为正的;(v)观测到的解释变量 $\mathbf{x}$ 是线性独立的,并有充分变化。各种不同的证明在这些条件上具有某种微妙的变化,不过,我们在这里将不探究这些问题。

非参数识别问题涉及数学上相当深奥的内容,此问题也与参数模型的内容有关。倘若人们设定参数形式诸如 $\lambda_0(t|\alpha)$、$\phi(\mathbf{x}|\beta)$、$g(\nu|\gamma)$,则已知数据时,这些函数会是唯一的吗? 不幸的是,在许多情况下,回答是"否定的"。这意味着一个研究人员可以估计一种特殊混合模型,而不考虑计算上的问题,却关注"良好"的结果与有意义的系数。不过,这种表示可能并不唯一。另一个研究人员,在不同参数假设下,可能会得出等价的良好结果,却具有不同意义。也就是说,观测到的生存函数可能与基准风险的其他选择及异质性分布相一致[兰开斯特(Lancaster, 1990,第4章)]。利用2.2节的术语,具有本质上不同政策含义的不同结构模型可能拥有同样的简化式。很明显,这对参数应用研究提出了一个课题。一个引人注目的求解方法是,选取风险与异质性的灵活参数形式,或者采用偏似然分析的半参数方法。本节将继续对这个问题进行讨论。

18.4 异质性分布设定

关于系数估计值对可供选择的异质性假设的敏感性问题,在文献里得到了广泛讨论。需要对下面两种看起来明显矛盾的主张加以辨别:

1. 对不可观测异质性的参数设定经常表现出一些任意性。这样的设定严重地扭曲了风险函数的推断。因此，参数形式上灵活设定或非参数设定是人们所希望的。参见赫克曼和辛格（Heckman and Singer，1984a）。

2. 倘若基准风险函数被正确设定，则不可观测异质性的参数设定就显得相对无关紧要。当风险函数的设定拿不准或者是不正确的时候，利用不同的异质性参数假设，则会产生数据边缘分布的各种不同估计。参见曼顿、斯托拉德和沃佩尔（Manton，Stallard，and Vaupel，1986）。

这两种观点之间的明显矛盾，可如下解决。对风险函数进行设定会影响到 $f(t)$ 分布的一阶矩，可是一旦假定异质性与观测到协变量不相关，异质性的设定就影响 $f(t)$ 分布的二阶矩。当风险函数得到正确设定时，异质性分布的主要影响就体现在估计量的相对有效性上。

18.4.1 *具有伽玛异质性的离散 PH*

前面的考察提出，具有任意风险函数的比例风险函数要使模型有吸引力，就将它与特定的异质性假设相结合。哈恩和豪斯曼（Han and Hausman，1990）与迈耶（Meyer，1990）都曾经将由 17.10 节探讨的离散比例风险模型与伽玛异质性假设结合起来。他们的研究表明，当基准风险不是参数化估计值时，对可供选择的 $g(\nu)$ 函数形式表现出很小的敏感性。

对于设定性，重新考虑包含异质性项的式（17.3）：

$$\varepsilon_i = \ln\left(\int \lambda_0(\tau)d\tau\right) - \mathbf{x}_i'\boldsymbol{\beta} - \nu_i$$

将它代入对数似然表达式（17.44）。异质性项需要通过积分去掉。哈恩和豪斯曼在伽玛异质性假设下给出闭形式表达式，并讨论当已知它们灵活风险设定时，对参数假设体现出相对较小的敏感性的研究成果。

18.4.2 *异质性的其他模型*

前面讨论强调，威布尔—伽玛模型具有闭形式这种方便的计算特性。

如果观测到的边缘分布的尾部比与伽玛或对数正态情况相一致的尾部要厚，那么人们可以考虑曼德布罗特（Mandelbrot）分布的稳定族成员。霍高（Hougaard，1986）提出嵌入伽玛与逆高斯族的非常一般族[也可参见贾吉娅（Jaggia，1991b）]。严格稳定分布服从下述条件：p 个独立实现值之和应该拥有标度因子乘以该分布的乘积的分布。霍高（Hougaard，2000，附录 3.3）给出了此类性质的一个概括。

尽管更加高度参数化的异质性分布看起来引人注目，因为它拥有较大的一般性，却产生了两类问题。一类问题是，可利用的数据不足以允许我们去识别或准确地估计参数。首先，在不试图进行估计时，这种情形经常被人们所忽略。

第二类问题是计算问题。若混合密度没有闭形式，则它就以积分形式出现。所得到的似然函数也是积分形式的一些项。进行估计就需要使用诸如数值积分或蒙特卡罗积分这类计算机密集数值方法，这些方法已在第 12 章讨论过。需要此类

估计方法的混合模型的例子是威布尔对数正态混合，其中，不可观测异质性服从对数正态分布。异质性模型基于模拟估计已由古里耶克斯和蒙福特(Gourieroux and Monfort，1991，1996)讨论过，可参考12.2节的例子。

18.5 离散异质性与潜类别分析

上面分析假定，不可观测异质性具有连续分布，同时关注该连续分布参数的估计。

一种可供选择的方法是，假定个体样本从有限个潜类别(**latent classes**)比如说 q 个构成总体中采样，而样本的每个元素可被看成来自这 q 个潜在子总体或层之一。这类模型分别称为有限混合模型、半参数异质性模型(**semiparametric heterogeneity model**)[赫克曼和辛格(Heckman and Singer，1984a)]以及潜类别模型(**latent class model**)[艾特肯和鲁宾(Aitlcen and Rubin，1985)]。该模型引人注目的特性是，它会导致灵活的参数分布。在持续期限模型中，赫克曼和辛格(Heckman and Singer，1984a)对此模型加以分析、倡导和应用。

虽然这些流行模型是在持续期限背景下阐述的，但为了突出在其他地方的应用，可使用一般性记号。例如，参见20.4节。

18.5.1 *有限混合模型*

考察下述两个成分的有限混合模型。倘若样本是来自两个子总体的具有 pdf $f_1(t|\mu_1(\mathbf{x}))$ 与 $f_2(t|\mu_2(\mathbf{x}))$ 的一种概率混合，则 $\pi f_1(\cdot)+(1-\pi)f_2(\cdot)$ 定义两种成分有限混合，其中，$0\leqslant\pi\leqslant1$。也就是说，观测值分别以概率 π 与 $1-\pi$ 从 $f_1(\cdot)$ 与 $f_2(\cdot)$ 进行采样。待估参数是 (π,μ_1,μ_2)。参数 π 可被处理成常值，或对 logit 函数进一步参数化。因而，$\pi=\exp(\lambda)/[1+\exp(\lambda)]$ 与 λ 依次利用可观测协变量进一步参数化。因此，我们考虑两种类型的个体，一类源自 $f_1(\cdot)$，另一类源自 $f_2(\cdot)$。沿着这些线索思考时，存在一个先验情况，例如，假若某个潜在特性可以通过这种方式分割样本总体。一种可供选择的解释是，密度的线性组合对 t 的观测分布给出了一个好的近似。

在原则上，对于具有三种或更多成分的可加混合的推广可直接进行，却受限于潜在成分识别性问题。本章稍后将进一步讨论该问题。因此，在经验应用中，假若成分拥有正常解释，这就非常有益。在最简单的水平上，我们将每个子总体考虑成"类型"，但在许多情况下，可能给出更多的解释信息[林赛(Lindsey，1995)]。

有限混合模型的另一种解释是，针对总体异质性的离散表述。假如总体由 m 种同质子总体组成，这样的子总体通常称为成分(**components**)。假定诸如威布尔模型或指数模的这类参数模型可应用于每一种成分。假如第 j 种成分是整个总体的一小部分 π_j，而 $\sum\pi_j=1$。

最后，将该问题系统表述如下：在下面所有例子中，不可观测异质性项的分布都具有无穷多个支撑点。如果连续混合分布 $g(\nu_j)$ 能通过含有有限 m 个支撑点的离散分布加以逼近，用 $\pi_j(j=1,\cdots,m)$ 表示有限 m 个支撑点的分布，那么边缘(混

合)分布是:

$$h(t_i \mid \mathbf{x}_i, \pi_j, \boldsymbol{\beta}) = \sum_{j=1}^{m} f(t_i \mid \mathbf{x}_i, \nu_j, \boldsymbol{\beta}) \pi_j(\nu_j) \tag{18.18}$$

其中,ν_j 表示估计支撑点,π_j 是相应的概率。赫克曼和辛格(Heckman and Singer, 1984a)在持续期限建模中考察了不可观测异质性的半参数表示。与之密切有关的研究工作是韦德尔等人(Wedel et al., 1993)做出的,这些研究对潜类别的解释令人满意。若混合分布 π_j 不受限于任何参数假设,则称混合模型为 t 的半参数混合模型。

对有限混合模型的估计,在已知成分数目或未知成分数目的条件下完成。若已知分数 π_j,则成分分布的极大似然估计能够被估计出来。更一般地讲,比例 π_j, $j=1,\cdots,m$ 是未知的,并且估计既涉及 π_j,又涉及成分参数。后一种情况下的极大似然估计量被称为非参数极大似然估计量(NPMLE)。这里的非参数成分是类别个数,但严格地讲,它是半参数方法,因为它融合了成分的参数模型。当成分个数未知时,如同通常情况,就会产生推断的某些微妙问题。详细内容参见 18.5.4 节。

引出有限混合类别的一种明显动机是,这是一种自然且简单的研究总体异质性的方法。在许多情况下,利用很少的潜类别个数,考虑不可观测异质性比如同 18.2 节中连续性"类型"的那种研究更为简单。

18.5.2 *潜类别的解释*

有限混合模型与潜类别分析(**latent class analysis**)有关[艾特肯和鲁宾(Aitkin and Rubin, 1985);韦德尔等(Wedel et al., 1993)]。设 $d_i=(d_{i1},\cdots,d_{im})$定义一个指示(虚拟)变量,使得 $d_{ij}=\mathbf{1}(\sum_j d_{ij}=1)$表示 t_i 是从第 j 个(潜在)组或类别采样的,$i=1,\cdots,N$。也就是说,每个观测值可被认为是来自 m 个潜在子总体、类别或者"类型"之一。在下面讨论中,我们假定模型是可识别的。

此模型设定$(t_i \mid d_i, \boldsymbol{\mu}, \pi)$服从独立分布,具有密度:

$$\sum_{j=1}^{m} d_{ij} f(t_i \mid \mu_j) = \sum_{j=1}^{m} f(t_i \mid \mu_j)^{d_{ij}} \tag{18.19}$$

其中,$\mu_j = \mu(\mathbf{x}_j, \boldsymbol{\beta}_j)$,$\boldsymbol{\mu} = (\mu_1, \cdots, \mu_m)$,而$(d_i \mid \boldsymbol{\mu}, \boldsymbol{\pi})$ 是 iid 的并服从多项式分布:

$$\sum_{j=1}^{m} \pi_j^{d_{ij}}, \ 0 < \pi_j < 1, \ \sum_{j=1}^{m} \pi_j = 1 \tag{18.20}$$

最后两个关系蕴含:

$$(t_i \mid \boldsymbol{\mu}, \boldsymbol{\pi}) \overset{iid}{\sim} \sum_{j=1}^{m} \pi_j^{d_{ij}} f_j(t \mid \mu_j)^{d_{ij}}$$

从而,得到似然函数:

$$\mathrm{L}(\boldsymbol{\beta}, \boldsymbol{\pi} \mid \mathbf{t}) = \prod_{i=1}^{N} \sum_{j=1}^{m} \pi_j^{d_{ij}} f_j(t; \mu_j)^{d_{ij}} \tag{18.21}$$

18.5.3 *EM 算法*

这种似然函数可直接求极大值,或利用 *EM* 算法求极大值,其中,变量 $\mathbf{d}=(d_1,\cdots,d_n)$被处理为缺失数据;参见 10.3 节。倘若 $\mathbf{d}$ 是可观测的,则模型的对数

似然是：

$$\ln L(\boldsymbol{\mu},\boldsymbol{\pi}\mid\mathbf{t},\mathbf{d})=\sum_{i=1}^{N}\sum_{j=1}^{m}d_{ij}\ln f_j(\mathbf{t}_i;\mu_j)+\sum_{i=1}^{N}\sum_{j=1}^{m}d_{ij}\ln\pi_j \qquad (18.22)$$

当已知 π_j 时，$j=1,\cdots,m$，则观测值 t_i 属于总体 j 的后验概率，$j=1,\cdots,m$，记为 z_{ij}，定义为：

$$z_{ij}\equiv\Pr[y_i\in 总体\ j]=\frac{\pi_j f_j(y_i\mid\mathbf{x}_i,\boldsymbol{\beta}_j)}{\sum_{j=1}^{m}\pi_j f_j(y_i\mid\mathbf{x}_i,\boldsymbol{\beta}_j)} \qquad (18.23)$$

z_{ij} 关于 i 的平均值是随机选取个体属于子总体 j 的概率。这等于 π_j，即：

$$\mathrm{E}[z_{ij}]=\pi_j$$

假定我们已知 $\mathrm{E}[d_{ij}]$ 的估计值 $\hat{z}_{ij}$。那么，以此估计值为条件，得到：

$$\mathrm{EL}(\boldsymbol{\beta}_1,\cdots,\boldsymbol{\beta}_m,\boldsymbol{\pi}\mid\mathbf{t},\hat{\mathbf{z}},\mathbf{x}_1,\cdots,\mathbf{x}_m)=\sum_{i=1}^{N}\sum_{j=1}^{m}\hat{z}_{ij}\ln f_j(t_i,\mu(\mathbf{x}_j,\boldsymbol{\beta}_j))+\sum_{i=1}^{N}\sum_{j=1}^{m}\hat{z}_{ij}\ln\pi_j \qquad (18.24)$$

它构成了 EM 算法的 E 步骤。该算法的 M 步骤可通过求解一阶条件

$$\hat{\pi}_j-N^{-1}\sum_{j=1}^{m}\hat{z}_{ij}=0,\quad j=1,\cdots,m \qquad (18.25)$$

$$\sum_{i=1}^{N}\sum_{j=1}^{m}\hat{z}_{ij}\ \frac{\partial\ln f_j(t_i\mid\boldsymbol{\beta}_j)}{\partial\boldsymbol{\beta}_j}=\mathbf{0} \qquad (18.26)$$

来求 EL 极大值。其次，我们能用式(18.23)得出 $\hat{z}_{ij}$ 的一些新值，并通过 E 步骤与 M 步骤进行迭代。

18.5.4 *选取潜类别数量*

第一个重要问题涉及对 m 的选取，即成分个数。通常没有先验理论指南，而且一般地讲，选取都是建立在实用基础上。由于待估参数个数是 $m\ \dim[\boldsymbol{\beta}]+m-1$，所以参数数目相当大。倘若对 $\boldsymbol{\beta}$ 的某些元素限制成相等的，则这个数目或许会减少。一种流行的方法是，允许截距变化，但对斜率参数约束成对不同组是相同的[如同式(18.18)]。不过，若允许所有参数随不同类别而变化，很明显，人们愿意鼓励保持 m 不变。甚至仅当截距被允许变化时，许多应用都使用 $m=2$。一种切合实际的策略是，以 $m=2$ 开始，然后利用诊断检验去检查模型拟合情况。当拟合不好时，要添加其他成分。当所添加成分不能真实地被辨别时，这样做就产生了问题。当类别间差异小时，就不必使用有限混合表示。最令人满意的情况是，对成分都具有解释。对于不同维数的模型之间进行选择时，要使用惩罚似然准则(AIC 或 BIC)，参见 8.5.1 节。由于存在参数边界假设问题，所以似然比检验并不合适。贝克和梅琳诺(Balcer and Melino, 2000)阐述了，蒙特卡罗戏剧性地揭示出过度参数化的潜在陷阱，此类模型为了避免错误设定，不论是持续期限还是异质性都可被灵活设定。

当模型被过度参数化时，人们就不能识别参数。该问题通过存在多个最优解或平坦似然曲面而清楚地显示出来。计算算法依赖于初始值而可能收敛到不同的点。

利用惩罚似然准则从竞争模型中挑选出的模型不一定把样本数据描述得很好。这只能借助于合适的拟合优度检验与模型诊断才会查明情况。从本质上讲，人们能将持续期限的实际分布与拟合分布加以比较，两者之间的显著差异揭示出，模型的系统成分不足以解释观测样本变异。一些可能性将在下一节考虑。

计算上的考虑

第二个问题是对计算机算法的选择。尽管 EM 算法在认识问题的计算结构方面非常有益，但在实际应用中经常显得很慢。一些作者发现，基于数值导数的牛顿—拉夫森算法会产生许多令人满意结果的例子。参见霍顿(Haughton, 1997)对可供选择方法给出的一个综述。不论运用什么算法，若组间差异很小，则似然曲面将趋于揭示几个局部极大值。无论如何，无法保证仅有唯一极大值。

所有有限混合模型在以下情况，即在倘若对子总体标号加以排列，则数据分布没有变化时，都是不可识别的。也就是说，把"第 1 个"成分重新标号成"第 2 个"成分，或者反过来，并不会产生差异。这种问题能借助于对 π_j 或 μ_j 设定成非递减的而得以处理。令人满意的是，成分标号具有某种行为解释。

有限混合模型的一个潜在局限性是，额外成分可以直接反映出离群值的存在。虽然这不一定是一件坏事，但有用的是，能识别出对一个或多个成分负责的处于外面的观测值。在这一点上，式(18.23)就有用。一旦实施估计后(**postestimation**)，能计算后验概率。对于离散值来说，这些概率关于一个成分将是大的，而关于其余成分则是小的。

18.6 存量抽样与流动抽样

在许多实际应用情况中，会出现下述问题：在可利用的两个或多个不同平均持续期限测量之间的关系是什么？由人口学知道，平均年龄与预期寿命范围之间有众所周知的差异。在房地产业中，提供销售的资产仍未卖出的平均时期与新增加销售资产在卖出之前的预期时期之间是否存在差异呢？第一个概念经常用于普遍讨论，而第二个问题可能更有重大意义。在经济学中，在由政府统计局发布的失业持续期限的不同测量之间存在着类似的问题。不可观测异质性的问题因为它附属于失业者的汇合及不断进入那种汇合之中，所以与这些讨论密切有关。有关这些问题的早期有影响的讨论之一，由萨伦特(Salant, 1977)给出。

为了具体起见，我们关注于熟知的事业持续期限的例子。一种测量失业者个体的失业经历的统计量是，平均中断持续期限(**average interrupted duration**, 记为 AID)，在许多国家，由统计局发布此值，它是那些当前失业者存量成员仍然处于失业的一个平均时期。它是预期流逝持续期限(**expected elapsed duration**)的估计值，即那些新失业个体期望保持失业的时期，经常称为完整失业时期的平均持续期限

(ACD)。在工作搜寻文献中,它起着显著的作用,而且是本章及前面几章探讨的核心内容。这是完全持续期限(**completed duration**)的期望长度的估计值。我们可将AID考虑成基于存量的测量,而将ACD考虑成基于流量的测量,前者类似于总体的平均年龄,而后者类似于期望寿命范围(**expected life span**)。

研究此类问题的一个适宜的统计工具是更新论(**renewal theory**)。具有常值强度参数的平稳泊松过程就是更新过程的一个例子。在时间区间 dt 之内的更新个数意指事件个数。持续期限是相邻事件发生之间(即更新)的时间。对于给定状态中的个体来说,向后递归时间(**backward recurrence time**)意指由于更新而流逝的持续期限,而向前递归时间(**forward recurrence time**)意指从当前状态到过渡的持续期限。时间区间 $(0,t]$ 上事件的期望个数记为 $\mathrm{E}[N(t)]$,称为更新函数(**renewal function**),其极限 $\lim_{dt\to 0}d\mathrm{E}[N(t)]/dt$ 就是更新强度(**renewal intensity**),它确定了ACD与平均向后递归时间之间的关系。在下文中,我们关注某些著名结果。

萨伦特(Salant, 1977)已经证明,风险率的异质性提供了AID与ACD之间差异的一种重要认识。他的图式表示法给出影响到计算平均值的两个关键因素的直觉图。在图18.1里,纵轴测量日期时间,而水平轴代表调查的日期。存量抽样(**stock sampling**)意指在调查时期对那些处于已知状态的个体存量进行抽样。与之相比,流量抽样(**flow sampling**)意指我们对在特定区间进入状态的那些个体进行抽样。实施中的时期长度已由图上的垂直轴所示。就图示而言,标出9个长度的实现值,其中4个(S6、S7、S8和S9)正处于调查日期之中,而另外5个(S1、S2、S3、S4以及S5)在12个月调查时期内都已完成。若用 u_j 表示第 j 个因调查处于实施抽样时期的长度,则对我们的例子来说,$\mathrm{AID}=1/4(\sum_j u_j)$。若用 t_i 表示第 i 个因调查而完成抽样时期的长度,则 $\mathrm{ACD}=1/5(\sum t_i)$ 。

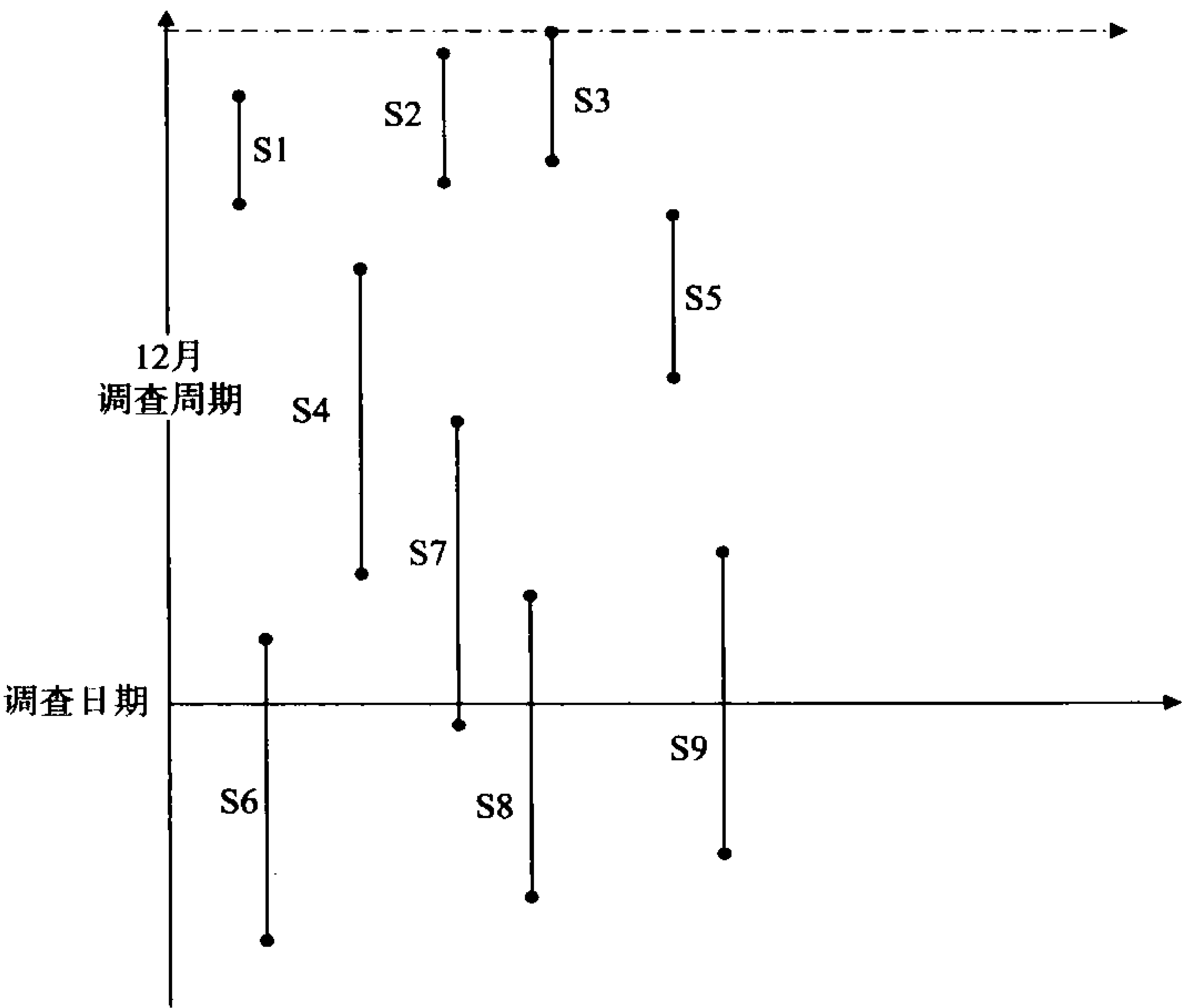


图 18.1 存量抽样下样本长度偏倚的例子

现在可以发现，与较短时期相比，调查更可能捕获到较长时期，而这会产生向上偏倚，也就是样本长度偏倚(**length-biased sampling**)。此类偏倚可能导致 AID>ACD。不过，由于调查仅测量了未完整持续期限，所以这种未完成持续期限的平均值可能小于完成持续期限的平均值。这就是中断偏倚(**interruption bias**)的现象。对哪一种偏倚占支配的问题的回答要依赖于时期长度分布，而且这反过来依赖于风险率的分布。异质性风险率提供了认识两者之间差异的重要内容。

一个重要假设是，平稳环境是指那种流入此状态与流出此状态均相等的情况。设 $f(u)$表示占用时期的密度，而 $g(t)$表示完成时期的密度。从而，u 的分布由：

$$f(u)=\frac{\bar{G}(u)}{\int\bar{G}(u)du}=\frac{\bar{G}(u)}{\mathrm{E}[t]} \tag{18.27}$$

给出，其中：

$$\bar{G}(u)=\int g(x)dx$$

表示对应于密度 $g(u)$ 的生存函数，而 $\mathrm{E}[t]$表示完成持续期限的分布均值。对于这个结果的全部推导与基本假设，参见萨伦特(Salant，1977)或兰开斯特(Lancaster，1990，5.3 节)。

该结果的意义在于，若 $g(t)$ 为指数的，事件的随机过程是泊松过程，则 $f(u)$ 也是指数的，并且 $g(t)$与 $f(u)$ 的持续期限均值相等。

已知式(18.27)，可以推导 u 分布与 t 分布之间的一般关系。一个有用结果是，将 u 的均值与 t 的均值及方差联系起来：

$$\mathrm{E}[u]=\frac{1}{2}\left(\mathrm{E}[t]+\frac{\mathrm{V}[t]}{\mathrm{E}[t]}\right) \tag{18.28}$$

另一个有意思的结果涉及 $\mathrm{E}[t]$与常值总体完整持续期限均值之间的关系，这里，常值总体具有正在实施的时期(也就是说，对不同的正在实施时期的存量加以平均)。按照基于样本长度偏倚抽样，此关系为：

$$\mathrm{E}[t^{(S)}]=\mathrm{E}[t]+\frac{\mathrm{V}[t]}{\mathrm{E}[t]}>\mathrm{E}[t] \tag{18.29}$$

这表明，常值存量的平均持续期限记为 $\mathrm{E}[t^{(S)}]$，$\mathrm{E}[t^{(S)}]$大于新时期的平均期望持续期限。若 $f(t)$是指数的，则 $\mathrm{E}[t^{(S)}]=2\mathrm{E}[t]$，并且 $\mathrm{E}[u]=1/2\mathrm{E}[t^{(S)}]$，一般来说，样本中断时期会部分完成。

倘若风险率不为常值，则会怎样呢？如果当风险率关于时期长度是递增的(即正状态相依性)，那么 $\mathrm{E}[u]<\mathrm{E}[t]$，而当风险率关于时期长度是递减的(即负状态相依性)，那么 $\mathrm{E}[u]>\mathrm{E}[t]$。

虽然这些结果是在常值总体假设下获得的，但它们在对各种通常所用的平均持续期限测量之间的联系加以解释与分类方面被证明非常有用。不论时期发生的原因如何，此处的结果都有效。这些结果还引发了对风险函数形状的更仔细严谨的探讨。

18.7　设定检验

在持续期限模型中，设定检验要用到几种不同形式，包括：

- 包含某些协变量与排除某些协变量；
- 对生存函数的函数形式检验；
- 不可观测异质性的检验；
- 状态相依性与不可观测异质性的联合检验。

第一种设定检验形式不会产生新的问题，并且借助于沃尔德形式检验加以解决。

倘若对不可观测异质性没有约束，则函数形式的约束检验与不可观测异质性的检验是相同的。因为后者使风险率的估计产生偏倚，如同 18.25 节所证明的，对不可观测异质性进行诊断检验是可取的。

对此而言，标准公式是检验异质性（方差）参数是否为 0。如果这个假设是利用假定零异质性的约束模型来加以检验，那么得分检验是合适的。若假设是一个边界假设，则使用基于无约束模型的似然比或沃尔德检验将会产生问题。例如，在威布尔—伽玛模型（18.9）中，约束 $1/\delta=0$ 将使模型特殊化成威布尔模型，但这是一个边界假设。在零假设下，一个标准自由度卡方检验服从加权的卡方分布。

18.7.1　*假设检验*

一种设定检验的形式是，建立在零假设下模型指数基础上的不可观测异质性的得分检验。由于异质性与持续期限相依性之间可能出现混淆，所以实行联合检验而非单独检验是令人满意的。这可利用局部异质性威布尔模型的框架来完成［兰开斯特（Lancaster，1985）］。

局部异质性密度（**locally heterogenous density**）通过考察任意密度在具有乘法异质性 v 的威布尔密度的 $\nu=1$ 处附近进行泰勒展开而产生，即：

$$
\begin{aligned}
S(t|\nu)&=e^{-\mu t^{\alpha}\nu}=e^{-\varepsilon\nu}\\
&=e^{-\varepsilon}[1+(-\varepsilon)(\nu-1)+(\varepsilon^2/2)(\nu-1)^2+O(\varepsilon^3)]
\end{aligned}
$$

其中 $\varepsilon=\mu t^{\alpha}$。由第二行：

$$
\mathrm{E}[e^{-\varepsilon\nu}]=e^{-\varepsilon}[1+(\varepsilon^2\sigma^2/2)]\equiv S_m(t)
$$

其中，σ^2 项表示异质性分布的方差。

于是，有：

$$
\begin{aligned}
f_m(t)&=-\frac{\partial S_m(t)}{\partial t}\\
&=\alpha\mu t^{\alpha-1}e^{-\varepsilon}[1+(\varepsilon^2\sigma^2/2)]-e^{-\varepsilon}[2\varepsilon(\alpha\mu t^{\alpha-1})\sigma^2/2]\\
&=\alpha\mu t^{\alpha-1}e^{-\varepsilon}[1+\sigma^2(\varepsilon^2-2\varepsilon)/2)]
\end{aligned}
$$

若利用上面结果，并考虑删失观测值，则对数似然是：

$$\begin{aligned}\ln \mathrm{L}(\alpha,\boldsymbol{\beta},\sigma^2) &= \sum_{i=1}^{N}\ln\{[f_m(t)]^{\delta_i}[S_m(t)]^{1-\delta_i}\}\\ &= \sum_{i=1}^{N}\delta_i[\ln\alpha+(\alpha-1)\ln t_i+\ln\mu_i+\ln(1+\sigma^2(\varepsilon_i^2-2\varepsilon_i)/2)-\varepsilon_i\\ &\quad+(1-\delta_i)\ln(1+\sigma^2\varepsilon_i^2/2)]\end{aligned}$$

其中，δ_i 表示删失指示变量，对于未删失持续期限，它取值为 1，否则取值为 0，$\ln\mu_i=\beta_0+\mathbf{x}_i'\boldsymbol{\beta}_i$，而 $\varepsilon_i=\mu_i t_i^{\alpha}$ 表示广义误差(**generalized error**)（参见 18.7.2 节）。

关注的零假设是 H_0：$\sigma^2=0$ 与 $\alpha=1$。这是一个零不可观测异质性与指数分布设定的联合检验。设 $\boldsymbol{\theta}=(\boldsymbol{\theta}_1',\boldsymbol{\theta}_2')$，$\boldsymbol{\theta}_1'=(\sigma^2,\alpha)$ 而 $\boldsymbol{\theta}_2'=(\beta_0,\boldsymbol{\beta}_1)$，并设 $\boldsymbol{\theta}_0'=(0,1,\beta_0,\boldsymbol{\beta}_1)$ 表示约束向量。

为了简便起见，只考察未删失数据情况。于是，联合得分检验统计量是：

$$\mathrm{LM_{HD}}=\frac{1}{d}\mathbf{s}'\begin{bmatrix}\Psi'(1) & 1\\ 1 & 1\end{bmatrix}\mathbf{s} \tag{18.30}$$

其中，$\mathbf{s}'=\left[\frac{1}{2}\sum_i(\varepsilon_i^2-2\varepsilon_i),\sum_i(1+(1-\varepsilon_i)\ln t_i)\right]$，而 $\Psi'(r)$ 表示双伽玛函数的一阶导数 $d\ln\Gamma(r)/dr$，并且 $d=1/(N(\Psi'(1)-1))$。为了进行检验，$\mathrm{LM_{HD}}$要在零假设下计算（也就是说，在指数分布零假设下用它们的估计值代替所有相应量）。此检验统计量服从渐近 $\chi^2(2)$分布[贾吉娅和特里维迪(Jaggia and Trivedi, 1994)]。

注意到，$\mathrm{LM_{HD}}$统计量的二次型矩阵不是对角的。也就是说，联合检验的两个成分是相关的。异质性（持续期限相依性）的单独检验针对持续期限相依性（异质性）来说是有效力的。更明确地讲，假设我们考虑异质性与持续期限的两个单独得分检验。它们分别是：

$$\mathrm{LM_H}=\frac{1}{4N}(\textstyle\sum_i(\varepsilon_i^2-2\varepsilon_i))^2 \tag{18.31}$$

$$\mathrm{LM_D}=\frac{1}{d}(\textstyle\sum_i(1+(1-\varepsilon_i)\ln t_i))^2 \tag{18.32}$$

其中每一个在零假设下，都服从 $\chi^2(1)$分布。零不可观测异质性的单独检验针对其他零假设来说是有效力的，这是因为检验是相关的，参见式(18.30)。因此，在单独检验基础上对错误设定方向进行推断可导致错误结论。

由于对不可观测异质性的设定与状态相依性是紧密联系的，所以对它们单独进行假设检验能够产生错误的结果[贾吉娅和特里维迪(Jaggia and Trivedi, 1994)]。更正式地讲，存在不正确忽略异质性条件下，对状态相依性的检验是有偏的，而且反之也是对的。贾吉娅(Jaggia, 1991c)重新分析了在经济计量学文献中以导致错误的方式被人们分析的罢工持续期限数据。贾吉娅和特里维迪(Jaggia and Trivedi, 1994)发展了参数模型中的某些联合检验。也可参见贝拉和龙恩(Bera and Yoon,1993)对当模型被错误设定时假设检验的更一般问题所进行的研究。

由于这些检验在简单参数模型中是有用的，所以研究的起点可以是威布尔模型、威布尔—伽玛模型或者比例风险模型。在此情况下，对不可观测异质性的检验

或任何其他设定误差，都能够利用综合风险函数来完成，这是因为在没有异质性条件下，综合风险是单位指数随机变量。现在，我们讨论评估基于综合风险的模型拟合的一些图形方法。

18.7.2 检测错误设定的图形工具

在 8.7.2 节，我们研究了广义残差的概念。在非线性模型中，要明确选择这类测量很困难。在当前背景下，存在一种好的选择。

广义残差

一种有用的检验形式是，对持续期限模型拟合进行非参数图形检验。该检验使用了广义残差，广义残差被定义成数据与待估计参数的某种函数。对于正确设定模型来说，残差应表现出大致像源自已知分布的 iid 样本一样。可以证明，综合风险具有这种性质，从而函数作为基于残差设定检验的成分。在 17.3.1 节的持续期限模型背景下，有：

$$
\begin{aligned}
S(t|\mu) &= \exp[-\Lambda(t|\mu)] \\
f(t|\mu) &= \lambda(t|\mu)\exp[-\Lambda(t|\mu)]
\end{aligned}
$$

考虑广义残差的分布：

$$
\begin{aligned}
\epsilon &= -\Lambda(t|\mu) \\
&= -\ln(S(t|\mu))
\end{aligned} \tag{18.33}
$$

这个变换的雅可比行列式是：

$$
\begin{aligned}
|J| &= dt/d\epsilon \\
&= \frac{1}{d\Lambda(t|\mu)/dt} \\
&= 1/\lambda(t|\mu)
\end{aligned}
$$

已知 $f(t|\mu)$，变换式(18.33)以及变换的雅可比行列式时，ϵ 的密度由

$$
\lambda(t|\mu)\exp(-\epsilon)\ \frac{1}{\lambda(t|\mu)} = \exp(-\epsilon) \tag{18.34}
$$

给出，它不依赖于 μ；此密度服从单位指数分布。这个结果可参考 17.3.1 节与 17.6.7 节。

基于综合风险的诊断检验

在正确设定零假设下，利用广义残差 ϵ 的单位指数性质可建立诊断检验。广义残差的生存函数是 $S(\epsilon)=\exp(-\epsilon)$。因此，$-\ln S(\epsilon)=\Lambda(\epsilon)=\epsilon$。对于正确设定模型来说，将估计综合风险与 $\hat{\epsilon}$ 进行图形比较，应该得到大致具有 45°斜率的正线性相关关系，如果点显著偏离 45°直线，则可能显示错误设定。

例如，威布尔模型的估计综合风险(**estimated integrated hazard**)是 $\hat{\epsilon}=\hat{\mu}\ t^{\hat{\alpha}}$。它的生存函数是 $\hat{S}(\hat{\epsilon})=N^{-1}$(样本观测值个数$\geqslant\hat{\epsilon}$)。

一种简单形式是，将 $-\ln S(\hat{\epsilon})$ 对 $\hat{\epsilon}$ 及截距进行回归，并检验截距是否为 0 且斜率是否为 1。

这种方法可应用于那种有可利用的综合风险表达式的任何参数模型。例如，威布尔—伽玛混合形式(通过令 $\alpha=1$，很容易专门化成指数—伽玛混合形式)的广义误差是 $\epsilon=k\ln[(k+\mu t^{\alpha})/k]$。为了应用此检验，已知($\mu,\alpha,k$)的估计值，计算 $\hat{\epsilon}$，然后画出 $\hat{\epsilon}$ 对 $-\ln\hat{S}(\hat{\epsilon})$ 的图形。

删失数据

在删失观测值的情况下，观测持续期限 $t=\min[T,L]$，其中，L 表示右删失限。当观测值大于 L 时，它就在 L 处进行删失。于是，广义误差 $\epsilon(t)$ 不服从单位指数分布。通过下述推导，得出一种对删失进行调整的建议关系：

$$\begin{aligned}\mathrm{E}[\epsilon(T)\mid T\geqslant L] &= \int_{\epsilon(L)}^{\infty}\frac{\epsilon f(\epsilon)}{S(\epsilon(L))}d\epsilon \\ &= \frac{1}{e^{-\epsilon(L)}}\left[\int_{\epsilon(L)}^{\infty}\epsilon e^{-\epsilon}d\epsilon\right] \\ &= \frac{1}{e^{-\epsilon(L)}}[1+\epsilon(L)e^{-\epsilon(L)}+e^{-\epsilon(L)}-1] \\ &= 1+\epsilon(L) \qquad (18.35)\end{aligned}$$

其中，用到了分部积分并且进行了简化。

这个关系建议，当数据未删失时，人们将广义误差估计成 $\hat{\epsilon}(t)=\hat{\varepsilon}(t)$；而当数据删失时，则将广义误差估计成 $\hat{\epsilon}(t)=1+\hat{\epsilon}(L)$。一些可以利用的结果表明，当删失比例不太大时，该方法在删失指数模型中发挥得非常好[贾吉娅和特里维迪(Jaggia and Trivedi)，1994；贾吉娅(Jaggia，1997)]。

18.7.3 *条件矩检验*

应用于广义残差的条件矩(**conditional moment**)框架(参见 8.2 节)，为设定检验提供了丰富方法。其思想可在对不可观测异质性进行检验的背景下得到阐明。

前面已经证明，综合风险函数是服从单位指数的随机变量，其均值为 1 且方差为 2。在此情况下，关注的条件二阶距约束是 $E[(\epsilon-1)]^2=\mathrm{V}[\epsilon]=1$，或者等价地：

$$\mathrm{E}[\epsilon^2-2]=0$$

而且，也可以产生较高矩约束，并且进行联合检验或单独检验。详细内容参见贾吉娅(Jaggia，1991a)。

18.8 不可观测异质性例子：失业持续期限

在本节，我们在存在不可观测异质性并用解析形式上易于处理的一种参数模型加以参数化的假设下，重新探讨 17.11 节的实证例子。

正如 18.7.2 节所讨论的，我们通过考察模型的估计拟合，使用图形工具来检查不可观测异质性存在的可能性。对于正确设定模型，其残差应该服从单位指数分布。人们通过计算与画出证实累积风险函数针对广义残差，非正式地评估模型拟合。对于正确设定模型，图形应该显示出斜率为 1 的近似直线。

图 18.2 与图 18.3 分别标明，没有异质性与有(伽玛)异质性的指数模型的广

义残差图形。正如我们从两个图形中看到的，模型的拟合在我们引入了不可观测异质性之后，仅仅进行了边缘改进。

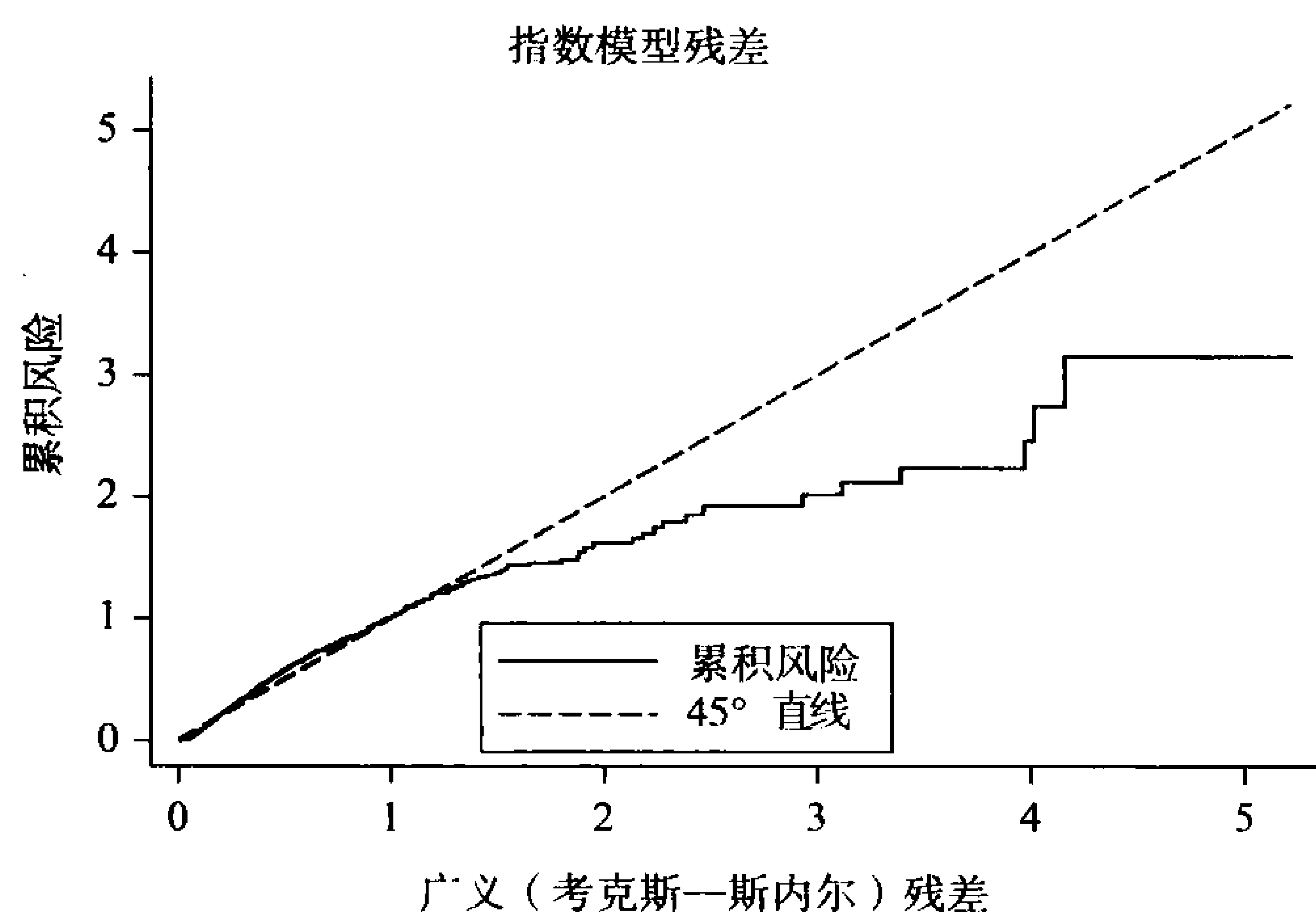


图 18.2 失业持续期限，来自指数模型的广义残差。美国数据为 1986～1992 年共 3 343 个时期，有些是完整时期。

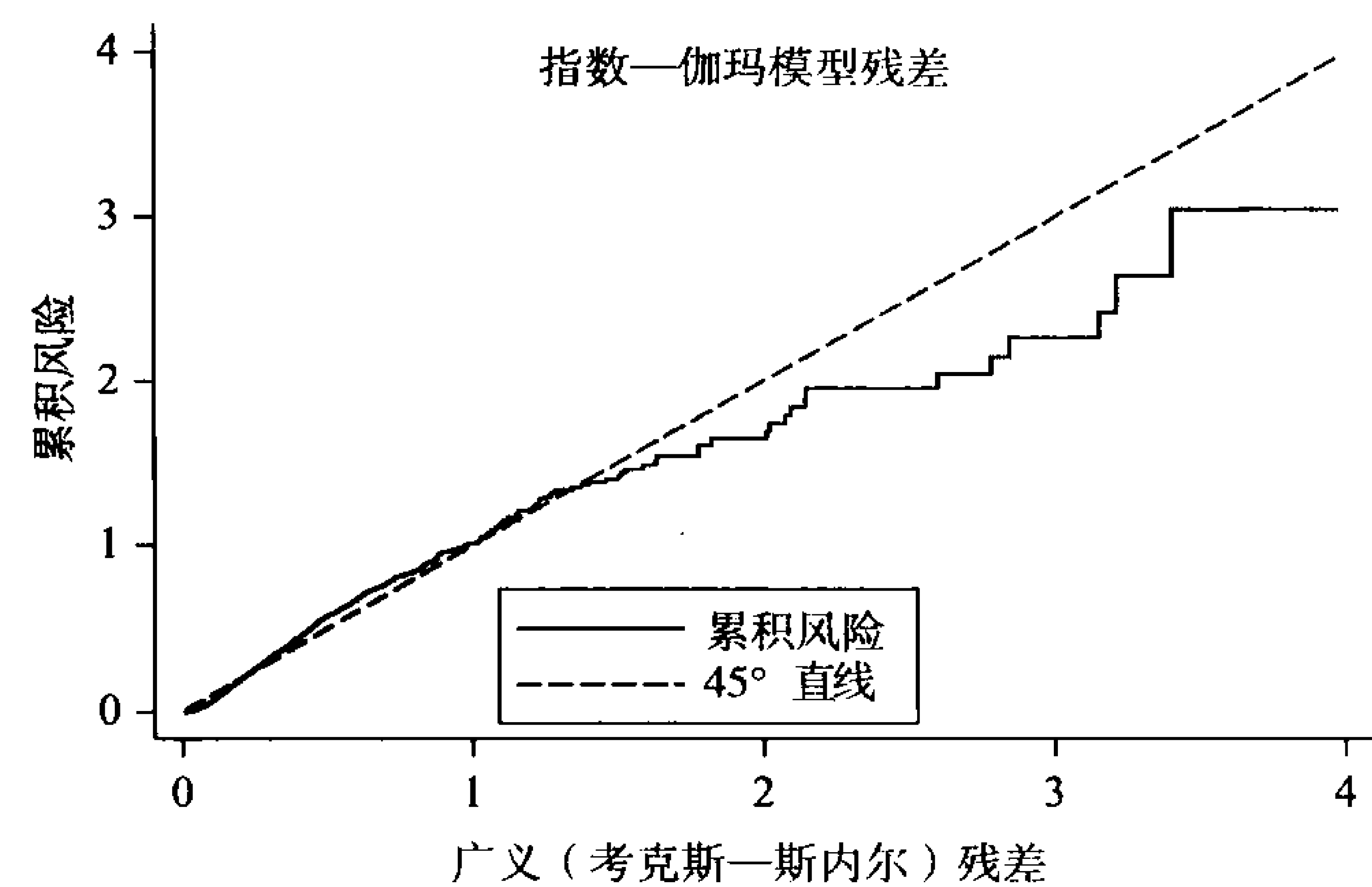


图 18.3 失业持续期限，来自指数伽玛模型的广义残差。数据与图 18.2 中的一样。

这个结果通过表 18.1 所列出的实际估计值得到证实，该表还展示了带有逆高斯异质性的指数模型的估计。尽管不可观测异质性存在显著证据，但在这两种背景下的系数估计值与我们先前在没有不可观测异质性条件下所获得的估计值并没有太大的差异。人们期望，不可观测异质性的存在将会对持续相依性参数有很大的影响，因为指数模型缺乏这个因素。

不过，当我们考察具有持续相依性与不可观测异质性时，会产生更有意思的情况。在没有假定它是"正确"模型的条件下，我们考虑威布尔分布逆高斯混合模型。为了方便比较，我们阐述表 18.2 中的这些估计，以及忽略不可观测异质性时的那些估计。

不可观测异质性的引入对持续期限参数有相当大的影响，它表现在从表 17.8

中的 1.128 到表 18.2 中的 1.753。后者蕴含着与忽略不可观测异质性的情况相比，离开失业的风险率会激增。回顾 18.2.4 节，在比例风险模型中，忽略异质性的结果之一是低估了风险率，因此，上述的发现证实与理论相符合。其次，注意到，不可观测异质性的证据极其强烈；估计方差参数 σ^2 具有大于 11 的 t 比率。再次，模型的拟合情况正如对数似然中所发现的，它也从－2 687.6 变为－2 616.6。很显然，系数估计值在性质上并没有太大的变化，可是当引入不可观测异质性时，人们能更正式地表明，显著系数(UI，LNWANG，CONS)的影响。

表 18.1　失业持续期限：带有伽玛与 IG 异质性的指数模型

	指数-伽玛		指数- IG	
变量	系数	t	系数	t
RR	0.501	0.817	0.504	0.821
DR	−0.882	−1.118	−0.807	1.032
UI	−1.585	−6.043	−1.545	−5.994
RRUI	1.091	1.725	1.057	1.686
DRUI	0.057	0.055	−0.013	−0.012
LNWANG	0.379	3.184	0.373	3.156
CONS	−4.095	−4.507	−4.097	−4.545
σ^2	0.232	3.178	0.207	2.925
−ln L	2 695.35		2 696.48	

表 18.2　失业持续期限：带有 IG 异质性与没有 IG 异质性的威布尔模型

	威布尔- IG		威布尔	
变量	系数	t	系数	t
RR	0.736	0.812	0.448	0.70
DR	−1.073	−0.933	−0.427	−0.53
UI	−2.575	−6.698	−1.496	−5.67
RRUI	1.734	1.857	1.105	1.57
DRUI	−0.061	−0.039	−0.299	−0.28
LNWANG	0.576	3.259	0.37	2.99
CONS	−5.303	−3.953	−4.358	−4.74
α	1.753	44.19	1.129	51.44
σ^2	6.377	11.149	—	—
−ln L	2 616.6		2 687.6	

尽管模型拟合得到了改进，但新混合模型还是能被错误设定。我们再次使用图形方法作为一种非正式的设定检验。图 18.4 与图 18.5 分别画出，威布尔模型具有不可观测异质性与没有不可观测异质性的广义残差。图形显示，混合模型尽管比指数- IG 模型更为一般，却看起来像是被错误设定。为了重申观点，虽然比较简单的模型既没有考虑到持续相依性，也没有顾及不可观测异质性，表明了很小的错误设定图形证据，可是“改进”的设定在前两个方面都推广了模型，但看起来却仍

像是错误设定,这种结果类似于贾吉娅(Jaggia, 1991c)。这种明显令人困惑的难题,可通过如下推理加以解决。用异质性与持续期限相依性的交互作用解释结果。威布尔模型假定单调风险。不过,麦考尔(McCall, 1996)利用同样数据提供了浴缸形状的风险函数更为合适的证据。他设定了一种多项式基准风险函数,比这里所用的单调函数更缺少约束性。因此,对我们的结果给出一种合理解释是,同时考虑不可观测异质性与持续相依性的模型,该方法与对这两种因素都忽略的模型相比,更加容易检查出错误设定。

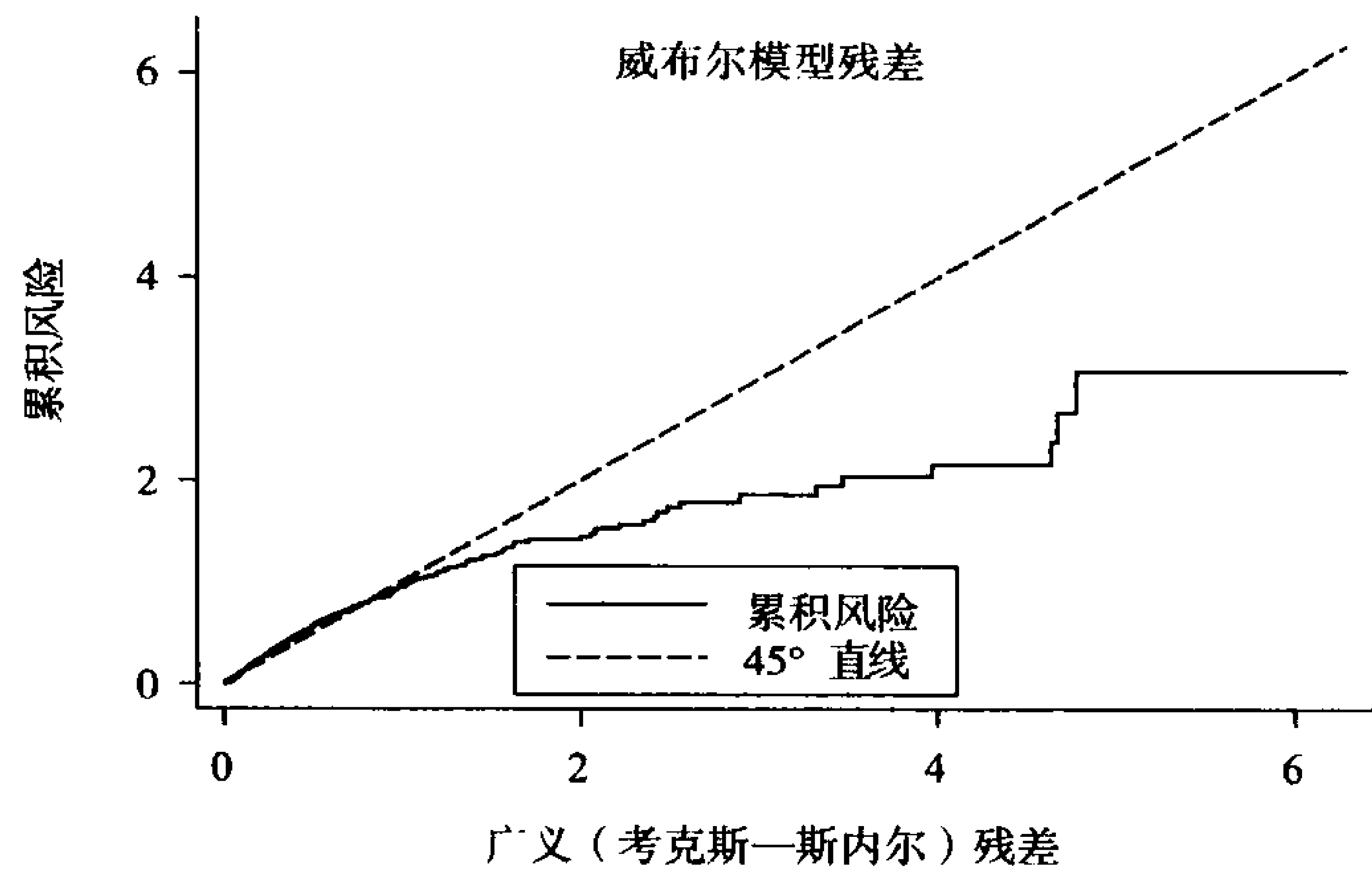


图 18.4 失业持续期限,来自威布尔模型广义残差。数据与图 18.2 的一样。

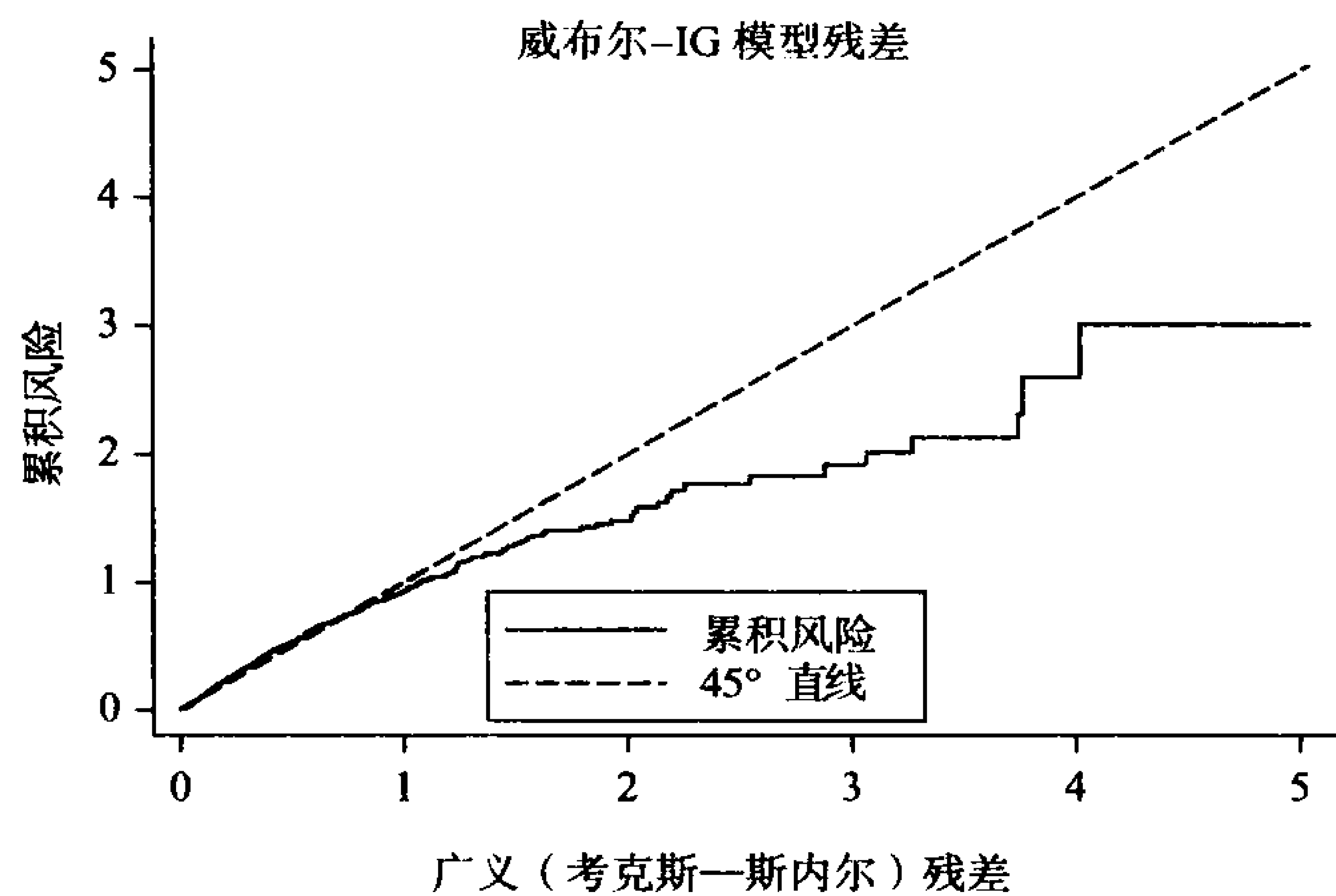


图 18.5 失业持续期限,来自威布尔—逆高斯模型的广义残差。数据与图 18.2 的一样。

最后,我们对不可观测异质性的存在与否进行参数检验。其目的在于阐明 18.7 节所讨论的理论。18.7 节发展起来的对忽略异质性的得分检验假定了未删失数据。由于这里所用的数据包括右删失观测值,故我们要实施由贾吉娅(Jaggia, 1997)发展起来的对删失样本的得分检验。

我们想要对指数持续期限模型中的零不可观测异质性 H_0: $\sigma^2=0$ 进行检验。设 $\boldsymbol{\theta}=(\sigma^2,\boldsymbol{\beta})$表示参数集合,并设 $\mathbf{s}(\boldsymbol{\theta}_0)$与 $\mathcal{I}(\boldsymbol{\theta}_0)$分别表示在零假设下计算的得分

与信息矩阵。利用18.7.1节推导的对数似然,可以写成 $\mathbf{s}(\boldsymbol{\theta}_0)=(\mathbf{s}_1(\boldsymbol{\theta}_0),\mathbf{s}_2(\boldsymbol{\theta}_0))$,其中,$\mathbf{s}_1(\boldsymbol{\theta}_0)=\frac{\partial\mathcal{L}}{\partial\sigma^2}\bigg|_{H_0}=\frac{1}{2}\sum(\epsilon_i^2-2C_i\epsilon_i)$,并且 $\mathcal{I}(\theta_0)=-\mathrm{E}\left[\frac{\partial^2\mathcal{L}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]\bigg|_{H_0}$。于是,不可观测异质性的得分检验是:

$$\mathrm{LM}=\mathbf{s}_1'(\tilde{\boldsymbol{\theta}}_0)\mathcal{I}^{11}(\tilde{\boldsymbol{\theta}}_0)\mathbf{s}_i(\tilde{\boldsymbol{\theta}}_0)\sim\chi^2(1) \tag{18.36}$$

其中,$\mathcal{I}^{11}=[\mathcal{I}_{11}-\mathcal{I}_{12}(\mathcal{I}_{22})^{-1}\mathcal{I}_{21}]^{-1}$ 表示贾吉娅(Jaggia, 1997)给出的 $\mathcal{I}(\boldsymbol{\theta})$ 分块逆对角成分,上标"~"用于表示约束极大似然估计值。

对于我们的样本来说,可以发现,LM=44.25,远远大于 $\chi^2(1)$ 的临界值,从而拒绝 $\sigma^2=0$ 零假设。该结果与来自威布尔—伽玛以及威布尔-IG模型的结果相一致,而后者因引入不可观测异质性,其模型拟合的显著性得到了改进。正如前面所关注的,这种检验针对错误设定持续期限相依性的检验有效力。

18.9 应用研究

风险函数与不可观测异质性之间交互作用的问题,已有大量文献进行研究。文献普遍认可的一种观点认为,若很好设定了风险函数,则异质性分布的精确参数设定相对而言并没有什么不妥[曼顿等人(Manton et al., 1986)]。该观点意味着,倘若风险函数被很好设定,而不是从参数形式上对不可观测异质性加以建模,我们就能简单运用稳健方差估计。其他一些研究建议,对异质性分布的参数设定并非无关紧要[赫克曼和辛格(Heckman and Singer, 1984a)],而使用非参数设定则是人们所盼望的。某种非常具有影响力的研究支持,运用含有非常灵活的风险函数设定,结合异质性的参数假设[迈耶(Meyer, 1990);哈恩和豪斯曼(Han and Hausman, 1990)]。最后,作为所有上述观点的一种折中情况,一些研究者运用哈恩—豪斯曼的离散时间方法或高阶多项式风险函数,并将它与非参数异质性的赫克曼—辛格方法相结合。不过,正如贝克和梅利诺(Baker and Melino, 2000)所指出的,这可能导致过度参数化,无疑将极为有害。因此,这看起来表现出对问题方法的敏感性,应小心谨慎对待此问题,宁愿运用简单的模型,而不用含有异质性参数的饱和模型。

考克斯PH模型,在生物计量学文献中占据中心位置。当人们对基准风险函数没有内生兴趣时,这看起来就是对函数形式的出色选择。建立模型从此处入手,常常是最好的。不过,不可观测异质性在大部分经济计量设定中极为重要,从而不应忽略它。

许多统计软件包经常提供能结合任何一种标准(伽玛、逆高斯或者对数正态)异质性(脆弱性)设定的标准参数持续期限模型的选择。尽管从使用上看,这是相当方便的,但是离散风险模型越是提供较大灵活性以及与经济数据的较好匹配,则越会表现出更大的吸引力。

对于潜在类别模型,运用EM算法经常遭遇到低的计算速度。通常,对似然函数直接求极大值既可行又有效。

18.10 文献注释

18.2 确实有许多论文已经讨论过异质性分布的设定及错误设定的后果。沃佩尔等人(Vaupel et al.，1979)给出了对伽玛模型性质的一个好的讨论。霍高(Hougaard，1995)给出异质性模型的一个综述。赫克曼和辛格(Heckman and Singer，1984a)曾提出非参数设定，并强调对错误设定的敏感性。曼顿等人(Manton et al.，1986)努力理顺对风险与异质性错误设定的相对重要性，建议前者是一个关键。

18.3 范登堡(Van den Berg，2001)提供了对 MPH 模型的透彻而易于理解的研究，以及有关 MPH 模型的设定方面的进一步参考文献。

18.4 哈恩和豪斯曼(Han and Hausman，1990)与迈耶(Meyer，1990)已经提出，灵活风险设定结合关于异质性的参数设定方面的一些好的实证例子。

18.5 赫克曼和辛格(Heckman and Singer，1984a)的论文是早期探讨离散异质性模型的。不可观测异质性的有限混合模型也被人们普遍称为“非参数异质性”模型。贝克和梅利诺(Baker and Melino，2000)阐述了持续期限和非参数异质性的蒙特卡罗研究。他们考察带有非参数异质性的持续期限相依性的非常灵活的设定。他们的结果表明，当这两者都存在时，在似然函数中拥有许多有限混合成分的策略会产生大的偏倚与不可靠的结论。利用 BIC 或汉纳—奎因(Hannan - Quinn)准则[1]即惩罚过渡参数化是有益的。

18.6 兰开斯特(Lancaster，1990)和萨特伦(Salant，1977)的书是关于样本长度偏倚的优秀参考书目。兰开斯特给出更新论的一个基础内容，几个重要结果构成了该理论基石。也可参见泰勒和卡林(Taylor and Karlin，1994)。

18.7 存在许多论文讨论了持续期限模型设定检验，其中的绝大多数研究针对比较易处理的无删失情况。基弗(Kiefer，1988)给出一个综述。贾吉娅(Jaggia，1997a)提出一个简短却清晰的关于设定检验的条件矩方法[格林(Greene，2003)也概括了此方法]。在持续期限模型条件下，目前尚未经检验的因计算需要而出现的设定检验则归功于安德鲁(Andrews，1997)。在计数模型条件下，卡梅伦和特里维迪(Cameron and Trivedi，1998，第 6 章)曾经讨论了有限混合模型的模型选择问题。基于持续期限模型的各种不同残差形式的模型诊断方面的一个好的介绍由

---

〔1〕 在统计学里，信息准则(HQC)是除 AIC 与贝叶斯准则(BIC)以外的另一种信息准则。HQC 可写成：

$$\mathrm{HQC}=n\ln\left(\frac{RSS}{n}\right)+2k\ln\ln(n)$$

其中，k 表示参数个数，n 表示观测值个数，而 RSS 表示来自线性回归或者非线性回归全局最优化的极小值的拟合残差平方和。

信息准则经常用作选择模型的指南。信息准则概念提供了在拟合优度与最少参数个数之间进行权衡的一种度量。从本质上看，构造信息准则统计量遵循的统计思想是一致的，即在考虑拟合残差的同时，依据自变量个数施加“惩罚”。不过，由此说它们是同一个指标确实不妥，毕竟“惩罚”力度还是不尽相同的。——译者注

霍斯默和乐梅肖(Hosmer and Lemeshow, 1999, 第196～200页)给出。

18.8 兰开斯特(Lancaster, 1979)的经典实证论文分析了,威布尔—伽玛混合模型条件下的失业持续期限。贾吉娅(Jaggia, 1991c)利用可以嵌入几种流行设定的广义伽玛模型,探讨了罢工持续期限模型中的错误设定。他的论文还突出了来自过度约束模型实施推断的困难。第19章涵盖持续期限的一系列其他应用。

习 题

18-1 [改编自萨普兰(Sapra,2002)]18.2节的分析证明了,不可观测异质性对无条件风险函数或平均风险函数的影响。强调忽略异质性会导致对平均风险函数的斜率低估。设条件风险函数是$\lambda_C(t|\nu)=\nu\lambda_0(t)$,其中,$\lambda_0$表示基准风险函数或无条件风险函数。证明:(i) 无条件风险$\lambda_U(t)<\lambda_0(t)$;(ii) 在下面每一个例子里,$\partial\lambda_U(t)/\partial t<0$。

(a) $\nu\sim\mathcal{U}[0, 1]$,并且$\lambda_0(t)=1, \ \forall t$。

(b) ν服从单位指数分布,满足pdf $g(\nu)=e^{-\nu}$,并且$\lambda_0(t)=\rho\exp(\gamma t), \ \rho>0, \gamma<0$。

18-2 当用异质性服从对数正态分布且均值为1的假设,代替异质性服从伽玛分布的假设时,重新考虑18.2.3节的威布尔—伽玛模型。

(a) 证明在此情况下,不可能获得无条件风险函数的解析表达式。

(b) 将无条件风险的积分表达式代入17.6.3节给出的对数似然函数中。利用12.4节的基于模拟极大似然法。说明如下估计算法。详述似然极大值化所包含的各种步骤。

18-3 考察指数—伽玛混合模型。这种模型是MPH模型的一种特殊情况。就指数模型而言,以乘法异质性因子ν为条件的生存函数是$S(t|\nu)=\exp(-\mu t\nu)$, $\lambda>0$。无条件生存函数是由平均生存函数给出的。利用ν的密度$g(\nu)$作为权重函数对异质性总体求平均,所以$S(t)=\int_0^{\infty}S(t|\nu)g(\nu)d\nu$。假定$\nu$服从(两参数)伽玛分布,满足$g(v)=\delta^k\nu^{k-1}\exp(-\delta\nu)/\Gamma(k)$。

(a) 已知伽玛异质性,证明$S(t)=(1+\mu t/\delta)^{-k}$。

(b) 推导无条件持续期限密度函数$f(t)$与无条件风险函数$\lambda(t)$的表达式。这些一般表达式可通过令ν在1处的均值进行专门研究;也就是说,设$k=\delta$,则会产生指数—伽玛混合模型。将这个混合模型分布的均值及方差,与最初的指数分布的那些均值及方差加以比较。

(c) 假如随机变量ν服从两点分布,使得以概率π取值ν_1,而以概率$(1-\pi)$取值ν_2。对无条件生存函数的设定来说,该假设含义是什么呢?请解释你的答案。

18-4 在不可观测异质性(有些计算机软件,也称不可观测异质性为脆弱性,还可能有子命令来设定它)服从伽玛分布的假设下,利用来自前一章实证问题的麦考尔数据集合的样本,重新估计那些过渡到全日制就业(CENSOR1=1)的威布尔模型。

(a) 运用 18.7.2 节中的广义残差,检验模型错误设定的假设。

(b) 新模型会显示出持续期限相依性的性质吗? 该新模型会对数据给出更好的拟合吗? 请参考不可观测异质性与持续期限相依性之间的交互作用,解释上述结果。

(c) 在对数正态异质性假设下,重新完成(a)部分问题。有关持续期限相依性的结果,会显著地不同于伽玛异质性的那些结果吗?

19 多重风险模型

19.1 引 论

本章研究几种不同的持续期限模型，从广泛意义上讲，可将它们解释成多变量模型，此类模型既涵盖平行过渡，又涵盖重复过渡。任何涉及一个以上指定状态的模型都可被看成是多变量模型，这是因为分析将包括两个以上持续期限的联合分布。

我们所考察的模型，通过各种方式产生，并应用于形形色色的数据类型。尽管存在差异，但为了组织方便，它们被归入本章。

具体起见，考察几个例子。源自劳动经济学的一个熟悉模型涉及从失业到就业的过渡，或脱离劳动力。第一种过渡能进一步被分成回到原来职业或到一个新职业。这两种指定状态是互不相交的。失业时期可能通过过渡回到两个指定状态的任何一个而终止。该例子的一种变形是，考察失业个体是找到一份全日制或兼职工作，还是继续成为失业者。因而，存在三种可能状态（指定状态）。第 17 章与第 18 章的一些模型已经研究，在两种状态之间进行过渡。人们仍然能运用两状态方法处理这类数据。例如，状态 1 代表全日制就业情况，而状态 0 代表其他任何状态。如同前面一样，这会涉及对风险率进行建模。不过，也能利用具有三种状态与两种过渡，因而用两种风险函数刻画这种情况，人们若对每一个指定状态进行设定，更一般地讲，存在众多失败类型，我们希望对从已知状态到任何一种失败类型的过渡进行建模。在本章，我们要将前面两章所研究的概念工具推广到处理多重风险（失败）或多变量持续期限模型上。

将一些重要问题表述如下。

1. 如何对协变量与各种失效类型之间的关系建模?

2. 如何在特定研究条件集合下，对失效类型之间的交互作用建模?

3. 已知某个失效类型"移动"或所有其他失效类型，对某些失效类型的失效率应如何估计?

多变量持续期限模型（**multivariate duration model**）涉及对所有过渡进行联立建模，即对两个或更多风险函数的联合设定与估计。分析多变量持续期限数据有几种可能框架；竞争风险（**competing risk**）框架就是最流行的一种框架。麦考尔（McCall，1996）提供了用于关注失业保险作用的失业数据的竞争风险框架的实证

应用。利用类似于麦考尔的方法，登格、奎格利和凡·奥德(Deng, Quigly and Van Order, 2000)探讨了抵押持有者采取抵押预先支付或抵押到期的过渡方式。

进行风险联合建模的动机是什么呢？同时，这样建模又会得到什么好处呢？若各种不同风险本质上是独立的，则各自分开建模与联合建模将会得出同样结果。不过，不同风险可能是有联系的；就每一种风险函数而言，可能存在一个共同的不可观测异质性项。否则，每一种风险可能包含具有一个或多个共同拥有成分的不可观测异质性，从而出现相关风险。

第二类例子包括人们分析指定状态持续期限的联合分布时，遇到平行事件的情况。比如，数对(T_1, T_2)代表失业持续期限与没有健康保险的持续期限。这里对风险进行联合估计的动机类似于前面所概述的情形。

第三个例子涉及在同一状态下重复时期(**repeat spells**)长度的联合分布(比如失业或没有健康保险)。即对给定个体来说，人们想要对终止时期的风险进行联立建模。如果所研究的时期是独立的，那么可利用前面几章的单时期方法对它们加以分析。若研究者希望探索过渡的相依性结构，则对已知状态的一些时期联立建模较为适宜。当出现时期相关时，需要新的模型及方法。上面的最后一个例子就比前面的例子更为复杂，因为由时间区间分开的事件之间可能存在相关性。例如，先前时期的长度与类型，或更一般地讲，时期的过去历史可能会影响到后续时期的概率与长度；或者个体的不可观测特征可能在后续时期持续。这类序列相关的不可观测异质性在重复时期之间产生了联系。正如一个事件的出现概率可能依赖于相同事件以前出现的情况。赫克曼和博尔哈斯(Heckman and Borjas, 1980)利用诸如出现相依性(**occurrence dependence**)与(马尔可夫)滞后持续期限相依性(**lagged duration dependence**)的概念，刻画个体状态相依性的几种结构类型。

文献中有大量模型对应于这些各种不同数据的情况。不过，尽管看起来模型有截然不同的选择，但通过一些共同线索将它们联系起来。在19.2节引入一些基本概念之后，我们研究流行的竞争风险模型。在19.3节，我们考察基于一组生存时间边缘分布的多元变量模型，并引入对生存时间进行联合建模的联接方法(**copula approach**)。多重时期建模，则在19.4节加以研究。

19.2 竞争风险

首先，我们引进经常用于竞争风险模型(**competing risks model**，记为CRM)和其他多变量公式的一些概念。这些概念常常是第17章曾引入概念的推广。当退出是一系列竞争状态时，基本竞争风险模型公式适用于对一个状态时间进行建模。竞争风险模型备受人们青睐，这是因为如果模型是PH模型，那么它相对可直接实施。

19.2.1 *基本概念*

现在，我们考察竞争风险模型，它有 m 个潜在持续期限或失效时间，其中每一种竞争都会引发失效。

潜在持续期限

对模型背景设置如下。每个实验者都具有基本失效时间，失效时间受限于删失。失效时间可能是 m 种不同类型之一，由集合 $J=\{1,\cdots,m\}$ 给出。我们将这个失效时间看成是过渡到已知状态（“死之”）的 m 种明显原因。不过，一类事件出现失效就会消除个体来自其他类事件的风险。因此，已知对每个个体的继续存在 $(m-1)$ 个持续期间的删失，我们就至多只能观察到一个完整持续期限。

具有 m 种失效类型的竞争风险模型，存在 $m+1$ 个状态 $\{0,1,\cdots,m\}$，其中，0 表示最初状态，而 $\{1,\cdots,m\}$ 是可能的指定状态。对于第 i 个个体，数据向量是 $(\mathbf{x}_i,t_i,d_{1i},\cdots,d_{mi},d_{ci})$ 形式，其中，$\mathbf{x}_i$ 表示测量 i 特征的弱外生协变量的向量，$t_i=\min(t_{1i},\cdots,t_{mi},t_{ci})$，其中，$t_{ki}$ 表示过渡到第 k 个指定状态的时间，t_{ci} 表示删失的时间，$d_{ji}\equiv\mathbf{1}(t_{ji}=t_i)$，$c$ 表示虚拟变量，当 $t_{ji}=t_i$ 时，c 取值为 1。由于我们唯一地观察到一个 t_{ji}，所以将其余的变量解释成潜在变量。

人们可将删失看成是一种竞争风险。根据概率分布，它对个体产生影响。在本章，删失变量被假定成与 $(t_1,\cdots,t_m)$ 是独立的。

i 的不可观测特性被纳入不可观测异质性中，用 ν 表示。当 ν 随退出原因而变化时，就将它写成 ν_j，$j=1,\cdots,m$。

竞争原因

竞争风险的一个标准例子是，由竞争原因而导致的死亡。考虑一位必须接受肾脏移植手术的个体，他处于过渡到健康状态、排斥状态或某种其他不健康状态的“风险”中，比如肝脏出现问题。若病死于任何一种状态，则意味着不可能过渡到其他状态。因此，在 m 种事件设置下，每一种事件都提供一种完整持续期限以及 $m-1$ 个删失持续期限。因而，我们拥有如下的“竞争风险”情况：确定该病人的指定状态时存在竞争。

尽管经验应用经常需要离散时间模型，可是对联合风险公式的解释却使用连续时间框架，而且一般来说，遵循米利和帕德尼(Mealli and Pudney, 1996)给出的解释。另外，我们假定拥有单时期数据。

模型提供时期持续期限(**spell duration**)的联合分布，记为 τ，而退出路线(**exit route**)为 r，r 是在集合 $(1,2,\cdots,m)$ 中取一个值的整数变量。

为了简单起见，我们忽略删失，并假定存在一些潜在变量 $(t_1,\cdots,t_m)$，每一种指定状态都具有一个潜在变量，倘若不存在其他风险因素引起该时期立刻结束的话，潜在变量通过哪一个时期可能结束对应于每个可能退出路线的时期持续期限。特定指定状态协变量用 $\mathbf{x}_j(j=1,\cdots,m)$ 表示。我们在时期终止时观测到一个持续期限 τ，其中：

$$\begin{aligned}\tau &=\min(t_1,\cdots,t_m) \\ &=\min_j(t_j),\quad t_j>0\end{aligned}\tag{19.1}$$

也就是说，只有最短持续期限是被观测到的，而其他持续期限都被删失了。这里，不考察归因于除退出以外其他因素引起的删失。于是，有：

$$\begin{aligned}\Pr[\tau>t]&=\Pr[t_1>t,\cdots,t_m>t] \\ &=S_\tau(t)\end{aligned}\tag{19.2}$$

这是一个联合生存函数。如果风险是独立的，那么：

$$\Pr[\tau>t]=\Pr[t_1>t]\times\Pr[t_2>t]\times\cdots\times\Pr[t_m>t] \tag{19.3}$$

相应的退出路线 r 由

$$r=\arg\min_{j\in J}(t_j) \tag{19.4}$$

给出。

若设 $g_j(t)dt$ 表示死于区间 $(t,t+dt)$ 内风险 j 的概率，则适用于所有原因的总风险率是：

$$\lambda_\tau(t)\equiv-d/dt\ln S_\tau(t)=\sum_{j=1}^{m}g_j(t)$$

在生物统计学中，这被称为总死亡力(**total force of mortality**)[戴维和莫什伯格(David and Moeschberger, 1978)]。如果风险是独立的，那么特定原因 j 的风险率是 $\lambda_j(t)=g_j(t)$。这意味着以一直生存到 t 为条件的位于 $(t,t+dt)$ 内的原因 j 引起的失效概率，关于 j 是众多风险之一还是唯一风险是相同的。

以一直生存到 T_1 为条件的位于区间 (T_1,T_2) 内的风险 j 的生存概率是：

$$\begin{aligned}\int_{T_1}^{T_2}\lambda_j(t)dt&=\int_0^{T_2}\lambda_j(t)dt-\int_0^{T_1}\lambda_j(t)dt\\&=\ln S(T_2)-\ln S(T_1)\\&=-\ln\frac{\Pr[t_j>T_2]}{\Pr[t_j>T_1]}\end{aligned} \tag{19.5}$$

或者等价地：

$$\exp\left(-\int_{T_1}^{T_2}\lambda_j(t)dt\right)=\frac{\Pr[t_j>T_2]}{\Pr[t_j>T_1]} \tag{19.6}$$

1 减去左边表达式称为位于区间 (T_1,T_2) 的原因 j 致死的净概率。表达式(19.6)有助于建立用于估计的似然函数。

独立风险

现在，我们明确地勾画出协变量影响风险率的图。假定独立风险(对应于相关风险)，并考察 t_j 的分布。第 j 种类型失效的风险率定义为：

$$\lambda_j(t_j\mid\mathbf{x}_j)=\lim_{\Delta t_j\to0}\frac{\Pr[t_j\leqslant T\leqslant t_j+\Delta t,\mid T\geqslant t_j,\mathbf{x}_j]}{\Delta t_j}$$

而第 j 种类型风险的综合风险率定义为：

$$\Lambda_j(t_j\mid\mathbf{x}_j)=\int_0^{t_j}\lambda_j(s\mid\mathbf{x}_j)ds$$

于是，若利用生存函数与综合风险函数之间的关系，则持续期限密度为：

$$\begin{aligned}f_j(t_j\mid\mathbf{x}_j,\boldsymbol{\beta}_j)&=\lambda_j(t_j\mid\mathbf{x}_j,\boldsymbol{\beta}_j)S_j(t_j\mid\mathbf{x}_j,\boldsymbol{\beta}_j)\\&=\lambda_j(t_j\mid\mathbf{x}_j,\boldsymbol{\beta}_j)\exp[-\Lambda_j(t_j\mid\mathbf{x}_j,\boldsymbol{\beta}_j)]\end{aligned}$$

一旦定义 $\mathbf{x}=[\mathbf{x}_1,\cdots,\mathbf{x}_m]'$，并且 $\boldsymbol{\beta}=[\boldsymbol{\beta}_1,\cdots,\boldsymbol{\beta}_m]'$，得出 τ 与 r 的联合密度：

$$f_j(\tau,r|\mathbf{x},\boldsymbol{\beta}) = f_r(\tau|\mathbf{x}_r,\boldsymbol{\beta}_r)\prod_{j\neq r}\exp[-\Lambda_j(\tau|\mathbf{x}_j,\boldsymbol{\beta}_j)] \tag{19.7}$$
$$= \lambda_r(\tau|\mathbf{x}_r,\boldsymbol{\beta}_r)\exp[-\Lambda_r(\tau|\mathbf{x}_r,\boldsymbol{\beta}_r)]\times\prod_{j\neq r}\exp[-\Lambda_j(\tau|\mathbf{x}_j,\boldsymbol{\beta}_j)]$$
$$= \lambda_r(\tau|\mathbf{x}_r,\boldsymbol{\beta}_r)\prod_{j=1}\exp[-\Lambda_j(\tau|\mathbf{x}_j,\boldsymbol{\beta}_j)]$$

第一行由条件概率与边缘概率的乘积得到。右边第二项则是除 r 之外所有退出路线生存概率的乘积,并用到了风险独立性假设。

式(19.7)蕴含:

$$\lambda_j(\tau|\mathbf{x}_j,\boldsymbol{\beta}_j)\exp\left[\sum_{j=1}^{m}-\Lambda_j(\tau|\mathbf{x}_j,\boldsymbol{\beta}_j)\right]=\lambda_j(\tau|\mathbf{x}_j,\boldsymbol{\beta}_j)\exp[-\Lambda^{a}(\tau|\mathbf{x},\boldsymbol{\beta})] \tag{19.8}$$

其中,$\Lambda^{a}(\tau|\mathbf{x},\boldsymbol{\beta})=\sum_{j=1}^{m}\Lambda_j(\tau|\mathbf{x}_j,\boldsymbol{\beta}_j)$表示总风险或整个综合风险。最后这个式子表明,脱离最初状态的总风险是所有指定状态的风险之和。整个生存函数是:

$$S(t)=\exp(-\Lambda^{a}(t)) \tag{19.9}$$

已知独立风险,似然函数是如同式(19.7)的所有观测值项的乘积。若所有函数形式都已设定,则这种似然函数就可以用显式形式写出。前面曾阐述的许多有意义的问题,比如函数形式的灵活性、不可观测异质性等,在 CRM 背景下仍是有意义的。与一般水平上的讨论相反,现在我们考察特定的函数形式。文献中的比例风险设定十分流行,这里将采用这样的设定。

19.2.2 *具有比例风险的 CRM*

这里的目标是推导时期长度的联合密度与退出理由,它通过对各种退出理由的综合风险进行加总而得到。

考察形式为:

$$\lambda_j(t;\mathbf{x})=\lambda_{0j}(t)\exp[\mathbf{x}'(t)\boldsymbol{\beta}_j],\quad j=1,\cdots,m$$

的 PH 模型,其中,基准风险 λ_{0j} 与 $\boldsymbol{\beta}_j$ 都具有类型 j 风险,$t_{j1}<\cdots<t_{jk_j}$ 表示类型 j 的 k_j 个有序失效。例如,当 $m=2$ 时,则 k_1 意指注册类型 1 失效的个体数目,而 k_2 意指注册类型 2 失效的个体数目。

就已知考克斯 CRM 而言,其似然函数是:

$$\mathrm{L}(\boldsymbol{\beta}_1,\cdots,\boldsymbol{\beta}_m)=\prod_{j=1}^{m}\prod_{i=1}^{k_j}\frac{\exp[\mathbf{x}_{ji}'(t_{ji})\boldsymbol{\beta}_j]}{\sum_{l\in R(t_{ji})}\exp[\mathbf{x}_l'(t_{ji})\boldsymbol{\beta}_j]} \tag{19.10}$$
$$=\prod_{j=1}^{m}\mathrm{L}_j(\boldsymbol{\beta}_j)$$

其中:

$$\mathrm{L}_j(\boldsymbol{\beta}_j)=\prod_{i=1}^{k_j}\frac{\exp[\mathbf{x}_{ji}'(t_{ji})\boldsymbol{\beta}_j]}{\sum_{l\in R(t_{ji})}\exp[\mathbf{x}_l'(t_{ji})\boldsymbol{\beta}_j]} \tag{19.11}$$

注意,这种似然函数具有下述四个特性:(1) $\mathrm{L}_j(\boldsymbol{\beta}_j)$ 是 17.8.2 节曾研究的偏

似然函数。由于基准风险函数不存在，故可应用前面所述的渐近结果。(2) 倘若风险是独立的，则对 $L(\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_m)$ 联立求极大值可通过使每一个个体因子 $L_j(\boldsymbol{\beta}_j)$ 极大化来得到；因此，不论是联立极大化还是各自极大化，其结果都是等价的。通过将标准渐近方法用于 m 项似然函数的每个个体因子，完成对各个 $\boldsymbol{\beta}_j$ 的估计与比较。(3) 将 17.7 节与 17.8 节的思想直接加以推广。若离散时间（虚拟变量）公式用于每一种风险类型，则对于具有 $\boldsymbol{\beta}_j$ 的每一种风险类型，风险函数的可识别成分能被联立估计出。(4) 不可观测异质性完全如同第 18 章单时期的两状态比较风险模型那样引入。

19.2.3 *CRM 的识别*

考克斯(Cox, 1962)与齐亚齐斯(Tsiatsis, 1975)证明了，当 CRM 没有协变量时，模型是不可识别的。更准确地讲，这意味着具有相关风险的任何 CRM 在观测形式上等价于具有独立风险的 CRM。不过，赫克曼和奥诺雷(Heckman and Honoré, 1989)已经证明，在某些假设下，具有混合 PH 形式并带有协变量的 CRM 是可识别的。范登堡(Van den Berg, 2001，第 3 438～3 441 页)提供了支持假设的一种解释。除第 17 章曾经讨论的那些假设之外，还需要一些假设。例如，协变量必须表现出“充分变异”，并且不应该是完全共线性的。此外，我们还需要，各个不同风险的基准风险不应该是完全相关的。

19.2.4 *回归系数的解释*

在 CRM 的比例风险形式公式里，协变量变动对那种来自已知状态的过渡风险率产生的影响类似于第 17 章的 PH 模型，可是，对回归系数进行直接解释所遇到的问题与 15.4.3 节对多项式 logit 讨论时遇到的解释问题相似。

不过，人们还可能对协变量变动对经由路线退出概率产生的影响感兴趣。这很难计算出来。为了理解这一点，注意到，经由路线 r 退出已知状态的概率的表达式由

$$\Pr[r \mid \tau, \mathbf{x}, \boldsymbol{\beta}] = \frac{\lambda_r(\tau \mid \mathbf{x}_r, \boldsymbol{\beta}_r)}{\sum_{j=1}^{m} \lambda_j(\tau \mid \mathbf{x}_j, \boldsymbol{\beta}_j)} \tag{19.12}$$

给出。由于协变量既出现在分子中又出现于分母中，而且分母是所有风险之和，所以偏导数 $\partial \Pr[r \mid \tau, \mathbf{x}, \boldsymbol{\beta}] / \partial x_{rk}$ 的符号依赖于模型中的所有参数。于是，β_{rk} 的符号也是此偏导数的符号，这一点并不成立（此处情形完全类似于第 15 章对多项式模型的讨论）。不过，倘若竞争风险具有比例风险形式，则可利用下述结果[托马斯(Thomas, 1996，第 31 页)]。当 $\beta_{rk} > \beta_{jk}$，$\forall j \neq r$ 时，$\partial \Pr[r \mid \tau, \mathbf{x}, \boldsymbol{\beta}] / \partial x_{rk}$ 的符号为正的。换句话说，当 $\lambda_r(\cdot)$ 的估计系数大于所有其他风险函数的相应系数时，则 x_k 增大导致经由路线 r 退出的条件概率增大。

19.2.5 *含有不可观测异质性的 CRM*

如果竞争风险具有比例风险形式，那么前一章的一些方法能够被推广到包括不可观测异质性的情况上。不可观测异质性的一般设定考虑到特定状态的随机成

分。设 $\boldsymbol{\nu}=(\nu_1,\cdots,\nu_m)$ 是乘法不可观测异质性项的向量,假定该异质性项具有联合分布函数 $G(\boldsymbol{\nu})$,那么:

$$
\begin{aligned}
f_j(\tau,r\mid\mathbf{x},\boldsymbol{\beta},\boldsymbol{\nu}) &= \lambda_j(\tau\mid\mathbf{x}_j,\boldsymbol{\beta}_j,\nu_j)\exp\Big[\sum_{j=1}^{m}-\Lambda_j(\tau\mid\mathbf{x}_j,\boldsymbol{\beta}_j,\nu_j)\Big] \\
&= \lambda_j(\tau\mid\mathbf{x}_j,\boldsymbol{\beta}_j)\nu_j\exp\Big[\sum_{j=1}^{m}-\Lambda_j(\tau\mid\mathbf{x}_j,\boldsymbol{\beta}_j)\nu_j\Big]
\end{aligned}
$$

其中,第二行是由乘法异质性的假设而得到的。

这是一个具有特定状态随机效应的竞争风险模型的例子。关于 $\boldsymbol{\nu}$ 的边缘分布,可通过针对 $\boldsymbol{\nu}$ 进行积分而获得:

$$
f_j(\tau,r\mid\mathbf{x},\boldsymbol{\beta}) = \int\cdots\int\lambda_j(\tau\mid\mathbf{x}_j,\boldsymbol{\beta}_j)\nu_j\exp\Big[\sum_{j=1}^{m}-\Lambda_j(\tau\mid\mathbf{x}_j,\boldsymbol{\beta}_j)\nu_j\Big]dG(\boldsymbol{\nu})
$$

上式包含一个 m 重积分。

一种可操控的情况是,$\boldsymbol{\nu}$ 的 m 个元素都是独立的并服从伽玛分布的随机变量。在此情况下,m 重积分分解成 m 个积分之积。一个例子是如下情况:对于每一个特定原因风险函数,我们具有威布尔—伽玛混合。在此情况下,竞争风险是独立的。

倘若我们允许 $\boldsymbol{\nu}$ 的元素成为相关的,则得到竞争风险是相关的一种更有趣的情况。实际上,这是一种广泛用于生成各个竞争风险之间相关性的“技巧”。具体地讲,假如我们有关于 $\boldsymbol{\nu}$ 的多变量对数正态分布,即 $[\ln\nu_1,\cdots,\ln\nu_m]'\sim\mathcal{N}[\mathbf{0},\boldsymbol{\Sigma}]$。这有两种后果。第一,它通过异质性而引起竞争风险的相关性;第二,它使得极大似然估计的计算变得相当困难。原因是后者作为一个 m 重积分没有解析表达式。因此,要应用蒙特卡罗积分。如果如同许多应用例子一样,m 等于 2 或 3,这仍是可操控的,但完全不是平凡情形。为了减少积分的维数,对协方差矩阵的结构加以约束可能会有用。例如,我们可使用因子结构,其中,每一项 ν_j 可能被设定为(比如说)两个 iid 随机变量的线性函数,这两个随机变量具有未知权重(因子载荷)。为了可识别性,可能必须对权重系数进行正规化约束。

19.2.6 含有相依竞争风险的 CRM

与那种通过各个不同竞争风险相关的异质性变量引起的相依性模型相比,独立 CRM 具有重要的计算优势。不过,后者会产生有关异质性结构的有价值的额外信息,诸如关联参数。然而,仍然存在着人们应该如何选择对相关异质性的设定加以约束的实际问题。为了便于解释,让我们考虑在像二元回归的设置条件下,运用下述类似于式(17.20)的设置:

$$
\begin{aligned}
\ln\Big[\int\lambda_1(u)du\Big] &= -\mathbf{x}'\boldsymbol{\beta}_1-\nu_1+\varepsilon \\
\ln\Big[\int\lambda_2(u)du\Big] &= -\mathbf{x}'\boldsymbol{\beta}_2-\nu_2+\varepsilon
\end{aligned}
$$

现在,我们能假定 $\nu_1=\nu_2=\nu$,也就是说,这两个风险模型有完全相同的不可观测异质性。该假设是,同样的不可观测因子都会影响时期,但它们的影响却不一样。这

相当于跨越两个风险完全相关的异质性。粗略地讲，我们能假定，比如 ν_1 与 ν_2 是相关的，并对关联参数进行估计。我们把这些分别考虑成异质性的单因子模型与两因子模型。从实证上看，更具约束性的方法是人们所希望的，这依赖于其内容。例如，若两个风险从属于一个相同个体，从而我们将 ν_1 与 ν_2 看成是反映特定个体因子，则单因子模型被证实为正确的。不过，倘若我们将两个因子看成是特定风险，则两因子模型更吸引人。当两因子模型是正确设定时，某种理论以及蒙特卡罗证据表明，运用单因子模型引起显著曲解[林德布姆和范登堡(Lindeboom and Van den Berg,1994)]。

19.3 联合持续期限分布

本节考察相交时期或者并行时期的情况，这里的时期是相关的。假设生存时间是连续的。解释是针对一般水平的，为了简单起见，要限制时期是未删失的且服从参数分布。

在联合分布生存时间的应用研究中，一个自然起点是，使用联合生存的或联合密度函数的特殊函数形式。存在可利用标准"函数形式"吗？或者，有一般方法用于生成前面几章曾讨论的模型的多变量对应内容吗？下面，我们就考察这些问题。

19.3.1 *多变量背景下的生存概念推广*

通过将前两章的定义与概率推广到多变量情况来开始是有益的。

一个多变量生存函数 $S(\mathbf{t})$ 被定义成：

$$S(\mathbf{t})=S(t_1,\cdots,t_q) \tag{19.13}$$
$$=\Pr[T_1>t_1,\cdots,T_q>t_q]$$

其中，$T_1,\cdots,T_q$ 表示 q 个生存时间，它们具有单变量生存函数 $S_j(t_j)$。由定义知：

$$S_j(t_j)=\Pr[T_j>t_j] \tag{19.14}$$
$$=S(T_1\geqslant 0,\cdots,T_j\geqslant t_j,\cdots,T_q\geqslant 0)$$
$$=S(0,\cdots,t_j,\cdots,0)$$

与单变量生存函数情况不同，有：

$$S(t_1,\cdots,t_q)\neq 1-F(t_1,\cdots,t_q)$$

例如，$S(t_1,t_2)=1-F(t_1)-F(t_2)+F(t_1,t_2)$。

$(t_1,\cdots,t_q)$的联合密度用 $f(t_1,\cdots,t_q)$表示；如果 $F(t_1,\cdots,t_q)$是连续的，那么：

$$f(t_1,\cdots,t_q)=(-1)^q\frac{\partial^q F(t_1,\cdots,t_q)}{\partial t_1\cdots\partial t_q} \tag{19.15}$$

与单变量情况相类似，联合风险函数是 $\lambda(t_1,\cdots,t_q)$，表示为：

$$\lambda(t_1,\cdots,t_q)=\frac{f(t_1,\cdots,t_q)}{S(t_1,\cdots,t_q)} \tag{19.16}$$

联合综合风险 $\Lambda(t_1,\cdots,t_q)$ 是 $\lambda(t_1,\cdots,t_q)$ 的 q 重积分。然而，$\Lambda(t_1,\cdots,t_q)$ 与 $S(t_1,\cdots,t_q)$ 之间并不存在着单变量情况的关系式。

已知这些定义，能推导出联合生存函数吗？克莱顿与库济克（Clayton and Cuziuk，1985）已经考察了二元变量模型，并阐述这里给出的定义。他们分析的起点是，关于“交叉风险比率”（**cross-hazard ratio**）函数的假设，该函数是给定 $T_2=t_2$ 与 $T_2\geqslant t_2$ 时，t_1 的两个条件风险函数。这就产生一个非线性的二阶偏微分方程，该方程的解生成一个联合生存函数，其中，交叉风险比率函数起着重要作用。我们详细阐述最初来源，但注意到，这种方法要求的一些假设可能很难被推广到比二元变量更多维数的情况。

19.3.2 *基于边缘的二元变量分布*

本节简要评述某些生成二元变量持续期限模型的方法。这种方法建立在有关边缘生存函数的假设基础上。倘若研究者对边缘分布有好感，并且希望用它们作为组成部分，这可能有益。当然，对组成部分的选择要对所得到的联合分布形式施加一些约束。

归功于马歇尔和奥利金（Marshall and Olkin，1990）的一种方法，以下述方式考察：两个失效时间的边缘分布中都含有乘法不可观测异质性的模型。设 $f_i(t_i\mid\mathbf{x}_i,\nu)$，$i=1,2$ 表示给定协变量 $\mathbf{x}_1,\mathbf{x}_2$ 时 t_1，t_2 的边缘分布，这里，ν 表示这两个边缘分布共同的不可观测异质性项，而且是两个风险之间关联的根源。在生存分析中，这类模型称为“共有脆弱性”模型；它是 t_1 与 t_2 之间相关性的（唯一）根源。假定 $\nu>0$，ν 服从密度为 $g(\nu)$ 的概率分布。t_1,t_2 的二变量分布被正式定义成：

$$f(t_1,t_2\mid\mathbf{x}_1,\mathbf{x}_2)=\int_0^{\infty}f_1(t_1\mid\mathbf{x}_1,\nu)f_2(t_2\mid\mathbf{x}_2,\nu)g(\nu)d\nu \tag{19.17}$$

其中，为了记号简单，没有使用分布参数。

这种生成为混合形式的二变量分布可有闭形式解或没有闭形式解。它也是所得到的二变量分布将 t_1 与 t_2 之间的相关关系限制成正的那种情况。在某些情况下，这可能并不是人们所希望的。

可用于任何数据类型的这种一般方法能被特定化成如下情况：用边缘生存函数（**marginal survivor functions**）代替边缘函数，然后对变量 ν 加以积分，推导联合生存函数（**joint survivor functions**）；因而，有：

$$S(t_1,t_2\mid\mathbf{x}_1,\mathbf{x}_2)=\int_0^{\infty}S_1(t_1\mid\mathbf{x}_1,\nu)S_2(t_2\mid\mathbf{x}_2,\nu)g(\nu)d\nu \tag{19.18}$$

应用这种思想的一个例子是由克莱顿和库济克（Clayton and Cuzick，1985）给出的，他们运用该公式在边缘比例风险含有伽玛异质性的假设下，获得二变量生存函数。

用于说明生成二变量生存模型的该方法有点受限制。其限制的一个根源是，单因子不可观测异质性的假设。原则上，此类限制容易去掉。例如，我们能用 (ν_1,ν_2) 代替 ν，其中，$\nu_1>0$ 和 $\nu_2>0$ 代表两个相关成分的向量，对生存函数都是具

体的，具有联合概率分布 $g(\nu_1,\nu_2)$。于是：

$$S(t_1,t_2 \mid \mathbf{x}_1,\mathbf{x}_2)=\int_0^{\infty}\int_0^{\infty}S_1(t_1 \mid \mathbf{x}_1,\nu_1)S_2(t_2 \mid \mathbf{x}_2,\nu_2)g(\nu_1,\nu_2)d\nu_1 d\nu_2 \qquad (19.19)$$

为了具体起见，假定：

$$\nu_1=\omega_{11}\varepsilon_1+\omega_{12}\varepsilon_2$$
$$\nu_2=\omega_{21}\varepsilon_1+\omega_{22}\varepsilon_2$$
$$\varepsilon_j \sim \mathcal{G}[1,\ \sigma_j^2],\quad j=1,2$$

其中，$\{\omega_{ij},\ i,j=1,2\}$是未知参数，经常被称为"因子载荷"(**factor loadings**)。这表明，当因子载荷不为 0 时，异质性成分(ν_1,ν_2)是 iid 的随机成分 ε_1 与 ε_2 的相关线性组合。在实证研究中，其他流行的假设是：(i) $(\ln\varepsilon_1,\ln\varepsilon_2)$服从标准二元正态分布；或(ii) ν_1、ν_2 服从离散(有限混合)分布。因此，模型(19.19)是一种二变量混合形式。另外的识别限制(比如，正规化 $\omega_{11}=1$)也是必不可少的。ν_1 与 ν_2 之间的皮尔逊相关系数，即 $\text{Cov}[\nu_1,\nu_2]/[V[\nu_1]V[\nu_2]]^{1/2}$，依赖于$\{\omega_{ij},\ \sigma_j^2,\ i,j=1,2\}$，而且可以直接证明，这个量并没有以通常的$-1$ 与$+1$ 作为下界与上界。(而且，注意到，对应的失效时间关联参数是 $\text{Cov}[t_1,t_2]/[V[t_1]V[t_2]]^{1/2}$，它确实不同于给定情况。)范登堡(Van den Berg，1997)针对具有常值基准风险的混合比例风险模型，推导出确切的 $\text{Cor}[t_1,t_2 \mid \mathbf{x}]$的上界，具体地讲，$-1/3<\text{Cor}[t_1,t_2 \mid \mathbf{x}]<1/2$，同时证明了，这些边界不依赖于协变量 x，也不依赖于异质性分布。另外，当基准风险不为常值时，相关性边界还是依赖于它们。

相对于那种不可观测异质性成分以未限制方式进入的情形，因子载荷设定具有计算优点，尽管单因子模型可能显得限制性太强，但未限制模型会产生潜在高维积分。从计算观点上看，所得到的分布可能容易处理，也可能不容易处理，这部分地依赖于积分是否会产生联合生存函数的闭形式表达式。倘若没有闭形式表达式，则需要用基于模拟的方法加以估计。目前，对此类模型进行估计已超出了标准软件包所涵盖的内容。

因子载荷设定对模型施加了一些约束[范登堡(Van den Berg，2001)，林登布姆和范登堡(Lindenboom and Van den Berg，1994)]。例如，如果一个边缘模型并不显示存在不可观测异质性，那么 $\text{Cov}[\nu_1,\nu_2]$一定为 0；当 $V[\nu_1]>0$ 且 $V[\nu_2]>0$ 时，$\text{Cov}[\nu_1,\nu_2]\neq 0$。因此，当 $\text{Cov}[\nu_1,\nu_2]=0$ 时，则边缘模型中的每一个都没有不可观测异质性。

从应用观点看，引人注目的多变量生存函数应该是灵活的。上面概述的方法存在一些局限性。人们已经提出了一些可供选择的方法。比较有把握的一种方法是，使用联接方法函数。霍高(Hougaard，2000，第 435～437 页)在生存分析背景下，对此种方法给出了一个介绍。

19.3.3 *联接方法*

联接(**coupla**)最初是由斯克拉(Sklar)于 1959 年在他的法文文章中[也可参见斯克拉(Sklar，1973)]引入的，当已知边缘分布，尤其是当人们想要以非正态分布

进行研究时，为了推导联合分布而提出的一种有用方法。虽然我们在联合生存模型背景下引进联接的思想，这里很容易应用它，但还可以用它探索离散、连续或者混合离散/连续型变量的任何集合的联合分布。

前面讨论过的一些方法（比如，马歇尔和奥利方法），通过不可观测异质性成分产生了变量间的相依性。在大部分应用中，这看起来引人注目，因为就观测到协变量而言，不可能涵盖经济事件的所有有关方面。

联接的性质

为了定义联接，我们以[0, 1]区间上的相依均匀随机变量 $U_1,\cdots,U_q$ 开始是可行的。相关关系[1]（**dependence relationship**），可通过随机变量的联合 cdf：

$$C(u_1,\cdots,u_q)=\Pr[U_1\leqslant u_1,\cdots,U_q\leqslant u_q] \tag{19.20}$$

加以描述，其中，函数 $C(\cdot)$ 表示联接，u_j 表示 U_j 的特殊实现值，$j=1,\cdots,q$。

右边是联合 cdf，即 $F(\cdot)$，而且联接的 q 个自变量能用 q 个边缘 cdf $F_1(\cdot),\cdots,F_q(\cdot)$代替。也就是说，联合 cdf 定义为：

$$C(F_1(u_1),\cdots,F_q(u_q))=F(u_1,\cdots,u_q)$$

就基于联接建立联合 cdf 而言，我们先选取一系列边缘，然后对所选边缘加以组合，以便生成联合 cdf。给定联接是关于所选边缘组合的函数形式，而对 $C(\cdot)$ 的不同选取会产生各种不同联合 cdf。斯克拉定理（**Sklar's theorem**）建立了下述内容；多变量分布函数能用式(19.20)的形式表示，同时若已知连续边缘，则该联接表示是唯一的。

当对多变量生存函数专门研究时，斯克拉定理表明，q 维多变量生存函数 $S(t_1,\cdots,t_q)$具有相应的联接表示 $C(S_1(t_1),\cdots,S_q(t_q))$。

考察 $q=2$ 的情况。于是，有：

$$\begin{aligned}F(t_1,t_2)&=\Pr[T_1\leqslant t_1,T_2\leqslant t_2]\\&=1-\Pr[T_1>t_1]-\Pr[T_2>t_2]+\Pr[T_1>t_1,T_2>t_2]\end{aligned}$$

而：

$$\begin{aligned}S(t_1,t_2)&=\Pr[T_1>t_1,T_2>t_2]\\&=1-F(t_1)-F(t_2)+F(t_1,t_2)\\&=S_1(t_1)+S_2(t_2)-1+C(1-S_1(t_1),1-S_2(t_2))\end{aligned}$$

其中，$C(\cdot)$称为生存联接（**survivor copula**）。现在注意到，$S(t_1,t_2)$仅仅是边缘生存函数的一种函数。

联接拥有下述某种对称性质：它允许以联接或生存联接开始研究[内尔逊(Nelsen，1999)]。乔(Joe, 1997)将与 $F(\cdot)$有关的二变量联接记为 $C(u,v)$，定义成单位正方形$[0,1]^2$ 上的二维概率分布函数。对于所有$(u,v)\in[0,1]$，$C(u,0)=C(0,v)=0$，$C(u,1)=u$ 并且 $C(1,v)=v$。在生存联接背景下，我们用边缘生存函数 $S(t_1)$代替 u，同时用第二个边缘生存函数 $S(t_2)$代替 v。在这种记号下，斯克拉

[1] 又称为相依性。——译者注

定理表明，存在一个联接函数 C，使得：

$$F(u,v)=C(F_u(u),F_v(v)) \tag{19.21}$$

其中，$F(u,v)=\Pr[U<u,V<v]$表示随机变量 U 与 V 的二元分函数，$F_u(u)$与 $F_v(v)$表示边缘分布函数。

若 F 为连续的，并且单变量边缘分布均有相应分位函数 F_u^{-1} 与 F_v^{-1}，则式(19.21)的唯一联接能被表述成：

$$C(u_1,u_2)=F(F_u^{-1}(u),F_v^{-1}(v))$$

联接方法涉及对每个随机变量的边缘分布进行设定，还要设定一个将它们连接起来的函数。对联接函数加以参数化，以便包括边缘分布之间的相关性测量。倘若没有检查出相关性，则这两个边缘分布是独立的，从而能分别对每一个变量加以估计。不过，若存在相关性，通过经由联接函数而重新得到的联合分布可求改进后的估计。由于无论边缘分布的形式怎样，联接都能获得相关性结构，所以有关变量建模的联接方法对经济计量学家来说，具有极为有用的潜在价值。弗雷谢界(**Frechet bounds**)使得借助于任何联接所容许探讨的相关性范围变得可行。

现在，考察具有 q 个持续期限($T_1,\cdots,T_q$)的一个例子，倘若有共同忽略不可观测异质性 ν，则 q 个持续期限是条件独立的；为了简单起见，将协变量剔除掉。于是，条件联合生存函数是：

$$\begin{aligned}\Pr[T_1>t_1,\cdots,T_q>t_q|\nu]&=\Pr[T_1>t_1|\nu]\times\cdots\times\Pr[T_q>t_q|\nu]\\&=S_1[(t_1)|\nu]\cdots S_q[(t_q)|\nu]\end{aligned}$$

并且多变量生存函数被定义成：

$$\Pr[T_1>t_1,\cdots,T_q>t_q]=\mathrm{E}_\nu[S_1(t_1)|\nu,\cdots,S_q(t_q)|\nu] \tag{19.22}$$

测算相关性

联接函数形式本身并不依赖于单变量边缘分布的形式。通常，联接是对能生成测量单变量边缘分布之间相关性的那种参数加以设定。一般地讲，相关性被参数化成一个纯量测量。为了简单起见，这里关注二变量联接。

对离散随机变量来说，联接表达式不一定是唯一的[乔(Joe,1997,第 14 页)]。在实际应用中，这不是一个主要问题，因为关心内容是去逼近一个未知的联合分布。建模关键问题是，选择联接函数的一个充分灵活的参数形式。

很难对来自联接的相关性参数给出一种解释，因为它们不一定位于[0,1]区间。因此，一种习惯做法是，将相关性参数(**dependence parameter**)转变成熟悉的关联性测量，比如肯德尔 τ 或斯皮尔曼 ρ；参见乔(Joe, 1997)。施魏策尔和沃尔夫(Schweizer and Wolff, 1981)已经证明，斯皮尔曼 ρ 相关系数只能根据联接函数加以表述；因而，有：

$$\rho(t_1,t_2)=12\iint\{C(u,v)-uv\}dudv$$

考察任何二变量联合 cdf $F(t_1,t_2)$，该联合 cdf 具有一元边缘 cdf $F_1(t_1)$与 cdf $F_2(t_2)$。由定义，$0\leqslant F_1(t_1),F_2(t_2)\leqslant 1$，这是因为每个边缘分布都在范围[0,1]

上取值。借助于弗雷谢下界 F^- 与上界 F^+，联合 cdf 是下有界的与上有界的，其中，F^- 与 F^+ 被定义成：

$$F(t_1,t_2)\geqslant F^-(t_1,t_2)\equiv\max[F_1(t_1)+F_2(t_2)-1,0]$$
$$F(t_1,t_2)\leqslant F^+(t_1,t_2)\equiv\min[F_1(t_1),F_2(t_2)]$$

由于联接是联合 cdf，故联接同样受限于弗雷谢界。弗雷谢界的信息在选择合适联接时极为重要。每一个联接都对其相关参数 θ 允许施加上界限。二变量联接的一个令人满意的性质是，当 θ 接近于其允许范围的下界（上界）时，该联接接近于弗雷谢下界（弗雷谢上界）。可是，一个联接的参数形式可能会加上一些约束，使得一个或两个弗雷谢界没有被包括在允许范围之内。因此，一个特定联接对某个数据集来说可能是更好的选择，但对另一个则不是。

例子

表 19.1 给出了文献中经常运用的某些二变量联接函数的一些例子。乔（Joe，1997）讨论了这些联接的性质。

表 19.1　某些标准联接函数

联接类型	函　数	定义域
乘积形式	uv	na[a]
FGMS[b]	$uv(1+\theta(1-u)(1-v))$	$-1<\theta<+1$
正态形式[c]	$\Phi[\Phi^{-1}(u)\Phi^{-1}(v);\theta]$	$-1<\theta<+1$
克莱顿	$(u^{-\theta}+v^{-\theta}-1)^{-1/\theta}$	$\theta\in(0,\infty)$
弗朗克	$-\theta^{-1}\ln(\eta-(1-e^{-\theta u})(1-e^{-\theta v}))/\eta,\ \eta=1-e^{-\theta}$	$\theta\in(-\infty,\infty)$

[a] na 表示不可应用。

[b] FGMS 表示 Farlie - Gumble - Morgenstern 联接。

[c] Φ 表示二变量正态 cdf。

正态联接与弗朗克（Frank）联接在其允许范围内包含两个弗雷谢界。克莱顿（Clayton）联接归属于阿基米德族（**Archimedean family**），其表达式为 $C(u,v)=\phi(\phi^{-1}(1-u)+\phi^{-1}(1-v))$；参见史密斯（Smith，2003）。

假如我们想要选择克莱顿联接，对二变量生存时间（t_1　t_2）进行建模。那么，依据边缘生存模型 $S(t_1)$ 与 $S(t_2)$ 表述的二变量分布将是：

$$(S(t_1)^{-\theta}+S(t_2)^{-\theta}-1)^{-1/\theta}$$

我们假定，边缘生存函数被设定成包括至多相差一个未知参数的形式。如同前面一样，可写出这些边缘生存函数，以便捕获到协变量与不可观测异质性的相关性。例如，将这些边缘生存函数建立在比例风险模型上。为了得到估计值，我们依据所得到的二变量联接，应用极大似然法。

这种方法并不是没有局限性的。特别地，有两点值得注意。第一，将该方法推广到三维或更高维情况并不容易。第二，人们不仅需要选择联接的特定函数形式，而且要认识到它在捕获给定数据集相关性方面的潜在限制性。例如，只支持正相关性。

推导来自联接的似然

为了拟合来自(以 cdf 定义的)联接的模型,第一步是选取一个联接,第二步是推导来自联接的(以 pdf 定义的)似然函数。一旦选定联接,就要考虑含有未删失失效时间(t_1,t_2)的二变量模型特殊情况的似然函数。定义$f_j(t_j)=\partial F_j(t_j)/\partial t_j$与$\partial C_j(F_1,F_2)/\partial t_j$,对于$j=1,2$,定义$C_{12}(F_1,F_2)=\partial C(F_1,F_2)/\partial t_1\partial t_2$。于是,概率密度为:

$$f(t_1,t_2)=f_1(t_1)f_2(t_2)C_{12}(F_1(t_1),F_2(t_2)) \tag{19.23}$$

其中,$f(t_1,t_2)=\partial^2F(t_1,t_2)/\partial t_1\partial t_2$,它用于构建其似然函数。若删失观测值出现在数据中,则必须适当修改似然函数。

运用各种不同联接可生成非嵌套模型。正如其他类似例子一样,惩罚对数似然(**penalized log-likelihood**)值能用于对各种联接的选取。

19.4 多重时期

本章前面引入平行状态[1](parallel states)与循环状态之间的差异是有益的。平行状态涉及一些平行事件,诸如处于就业与拥有健康保险;循环状态涉及序贯事件,比如第一次分娩、第二次分娩等。多重时期术语意指,同样事件的循环状态之间的持续期限。这类数据的联合建模类似于平行状态的联合建模,因为两者都涉及多变量概念,可是由于序贯事件可生成风险的动态相关性,故两者也有重要的差别。

考察一些循环事件的例子。劳动力市场中的个体者可能经历了就业与失业间的一系列过渡。例如,青年工人可能记录着一系列的失业时期。纽曼和麦卡洛克(Newman and McCulloch, 1984)考察了风险框架下的分娩时间。如果一个人想要对一系列分娩中的每一次分娩风险率进行建模,那么研究就必须给出分娩持续期限之间的相关性。特里维迪和亚历山大(Trivedi and Alexander, 1989)针对澳大利亚的青年人失业多重时期加以分析。在生育力文献中,连续分娩之间的持续期限是人们关注的内容[赫克曼、霍茨和沃克(Heckman, Hotz, and Walker, 1985)]。米利和帕德尼(Mealli and Pudney, 1996)运用英国退休调查数据,分析了就业与领取养老金情形持续期限之间的正相关性。恩格尔和拉塞尔(Engle and Russell, 1998)研究了股票市场上交易的特殊股票的连续交易之间持续期限的时间序列。史蒂文斯(Stevens, 1999)借助贫困的多重时期,分析了个体寿命中贫困的持续性。

上面提及的例子具有几个值得注意的特性。以先前事件为条件的事件风险率是否依赖于先前事件,这是一个重要的建模问题;第二,相关性的形式是人们关注的内容。先前时期的持续期限可能在确定后面事件的风险时,进入到协变量之中;先前事件的出现会影响到后面时期的基准风险。最后,不可观测异质性显示出序列相关性。上述每一个问题都是重要的建模问题。

多重时期(**mutliple spells**)生成了纵向数据或面板数据,这类数据会潜在地有

[1] 又称为并行状态。——译者注

助于解决如下重要识别问题：相对于风险函数中的异质性而言，动态相关性的影响所引致的识别问题。在某些假设下，多重观测值会使控制异质性更为容易，并且进行有关动态相关性的推断。

一般地讲，如人们所料，含有不可观测异质性与时期之间相关性的生存模型很难进行估计。不过，多重时期数据却创造了研究唯有利用面板数据才能探讨问题的机会。出现相关性、滞后持续期限相关性以及序列相关的不可观测异质性就是一些例子。不论出现滞后持续期限还是出现相关性，都意指正在研究中的先前时期个数或持续期限的终止概率的相关性。已知此类相关性，倘若忽略其相关性，则不适宜单独对时期加以研究。

考虑到为多重时期选择合适的经济计量框架，如同上一节所讨论的一样，一种可能性是运用联合生存函数对相关性建模。这种方法照顾到了数据的多变量特性。第二种可能性是，在没有忽略日历时间仍有关联可能时，使用面板数据框架，用时期下标代替时间下标。时期相关性会引发一些问题，这将在 22.5 节与 23.6 节的动态面板模型标题下加以讨论。在这两种情况下，由于面板损耗或者大部分最近时期的不完整而出现的删失可能性导致了重要差异。

19.4.1 *两时期模型*

运用两时期比例风险模型，可以阐明多重时期模型的一系列特性。在经济计量学中，此类模型已由奥诺雷（Honoré，1993）与霍罗威茨和李（Horowitz and Lee，2003）分析讨论过。

奥诺雷（Honoré，1993）曾经考察了形式为：

$$\lambda_s(t \mid \mathbf{x}, \nu) = \lambda_{0,s}(t)\phi(\mathbf{x}, \boldsymbol{\beta})\nu, \quad s = 1,2 \tag{19.24}$$

的比例风险模型。注意到，在该模型设定中，基准风险是特定时期的，可是异质性成分却不是，这里的该异质性以乘法形式进入表达式（一个重要假设），也就是说，ν 代表个体的固定或持久特征，从而得到一个固定效应模型。在类似于第 18 章讨论的混合 PH 那些条件下，他已经证明，该模型是可识别的。他还证明，对于识别来说，有关 ν 分布的假设不是基本的，协变量的存在也不是基本的。

在第二种模型中，奥诺雷考察了特定时期乘法异质性成分 ν_1 与 ν_2，ν_1 与 ν_2 具有联合二变量 pdf $g(\nu_1, \nu_2)$。ν_1 与 ν_2 之间的相关性反映出序列相关的异质性。这是一个随机效应模型。如同式（19.19）一样，利用混合分布 $g(\nu_1, \nu_2)$，借助于二变量混合方法可推导联合生存函数 $S(t_1, t_2 \mid \mathbf{x})$。若边缘生存函数是可识别的，则联合生存函数也是可识别的。该识别条件本质上是 PH 模型可识别性的那些条件。

奥诺雷还讨论了，两时期模型的滞后持续期限相关性设定，在第一个时期的持续期限（记为 t_1）以乘法形式进入第二个时期风险的假设中。他已经给出了，已知协变量与 t_1 时第二个时期条件模型的参数可识别性的充分条件。这里，就不讨论这些条件。不过，在这些条件下，比例风险模型的多重时期形式具有如下形式：

$$\begin{aligned} \lambda_1(t_1 \mid \mathbf{x}_1, \nu_1) &= \lambda_{0,1}(t)\phi(\mathbf{x}_1, \boldsymbol{\beta}_1)\nu_1 \\ \lambda_2(t_2 \mid \mathbf{x}_2, \nu_2) &= \lambda_{0,2}(t)\phi(\mathbf{x}_2^a, \boldsymbol{\beta}_2)\nu_2 \end{aligned} \tag{19.25}$$

其中，$\mathbf{x}_2^a=(\mathbf{x}_2,t_1)$表示协变量的增广向量。注意，当 ν_1 与 ν_2 相关时，这是一个内生性问题，在那种情况下，t_1 与 ν_2 不能是独立的。

先前出现的事件可能不会直接转变到后续时期的风险函数。通过引入新的协变量，也可能改进风险的设定。例如，失业时期可能导致对培训项目的注册，这似乎能影响到后面失业时期的风险。若将培训变量处理成弱外生的，则该模型的识别性受到威胁。这一点甚至与单一时期的模型分析有关：协变量与不可观测异质性是不相关的假设是有害的。

在一些情况下，人们可能愿意不仅对处于一个状态的多重时期进行建模，而且对于其他状态的那些时期也要加以建模。例如，存在两种状态，要么就业要么失业，我们不仅对最近失业时期的长度如何影响到当前失业时期感兴趣，而且对于预就业时期对摆脱失业风险的影响感兴趣。另外，当个体处于一个状态而不是另一个状态时，我们就能观测到个体的信息数据。例如，管理性数据只涵盖享受福利救助的人们，而没有涉及无福利求助的任何情况。

19.4.2 *更一般的多重时期模型*

为了阐明多重时期模型潜在计算复杂性，我们通过简略描述米利和帕德尼(Mealli and Pudney, 1996)模型开始讨论。

设 $\boldsymbol{\tau}=(\tau_1,\cdots,\tau_k)$表示 k 维完整时期向量，这里最初状态的指标为 r_{k-1} 的，而指定状态的指标为 r_k。假定在控制可能滞后持续期限相关之后，各个不同时期的持续期限是独立的。设 $\lambda_j(\mathbf{x}_j,\boldsymbol{\beta}_j)$表示特定指定风险函数，并设 $\mathbf{x}=[\mathbf{x}_1,\cdots,\mathbf{x}_k]$，$\boldsymbol{\beta}=[\boldsymbol{\beta}_1,\cdots,\boldsymbol{\beta}_k]$。

时期的联合密度及退出路线为：

$$
\begin{aligned}
&f(\tau_1,r_1,\tau_2,r_2,\cdots,\tau_k\mid\mathbf{x}_1,\cdots,\mathbf{x}_k,r_0,\boldsymbol{\beta}) \qquad (19.26)\\
&=f(\tau_1,r_1\mid\mathbf{x}_1,r_0;\boldsymbol{\beta})\cdots f(\tau_{k-1},r_{k-1}\mid\mathbf{x}_{k-1},r_0,r_1,\cdots,r_{k-2},\boldsymbol{\beta})\\
&\quad\times S(\tau_k\mid\mathbf{x}_k,r_0,r_1,\cdots,r_{k-1},\boldsymbol{\beta})\\
&=\prod_{j=1}^{k-1}\lambda_{r_j}(\tau_j\mid\mathbf{x}_j,\boldsymbol{\beta}_{r_j})\exp\Big(-\sum_{l=1}^{k}\Lambda_0(\tau_l\mid\mathbf{x}_l,\boldsymbol{\beta})\Big)
\end{aligned}
$$

这里，假定第 k 个时期是删失的(正在研究的)，并运用关系式(17.4)与式(17.6)。协变量包括随各个时期而变化的以及可能滞后持续期限。这个公式可与单时期 CRM 公式(19.7)相比较。

米利和帕德尼(Mealli and Pudney, 1996)运用该公式作为基础，建立一个精致模型。由于他们考虑到带有甚至比本章所探讨的更为复杂结构的不可观测异质性，所以其计算方法也更为复杂。他们使用了模拟极大似然法(参见 12.4 节)。

19.5 竞争风险例子：失业持续期限

第 17 章与第 18 章讨论的持续期限例子，关注于失业时期的时间，而忽略了过渡后的指定状态。这里，我们对麦考尔(McCall, 1996)所用的数据进行竞争风险分析。该数据区分了三种不同指定状态：在调查期间就业分别处于第 1 次安置后

全日制工作、第 1 次安置后兼职工作、第 1 次安置后全日制工作或兼职工作。因而，人们对这些假设放松如下：风险函数不依赖于指定状态，同时转而考虑竞争风险公式，其中，独立竞争风险决定了失业持续期限。

就麦考尔数据集而言，前面提及的三种状态分别有 1 073、339 以及 574 个过渡。第三种指定状态由于缺乏清晰解释，故对那种情况的结果不做详细讨论。对每一种过渡，我们都估计出四种参数持续期限模型，即含有逆高斯异质性的指数模型与威布尔模型，以及没有逆高斯异质性的指数模型与威布尔模型。尽管也可以考察伽玛异质性，但这类模型在计算上不稳定。由独立竞争风险的独立性假设，每一次估计一个方程。节选的计算机输出部分，已由表 19.2 与表 19.3 给出，这里仅关注第 17 章与第 18 章中的有限多个变量。

表 19.2　失业持续期限：含有 IG 脆弱性与没有 IG 脆弱性的指数模型的竞争风险估计值和独立风险估计值

	没有异质性			IG 异质性		
风险系数	风险 1	风险 2	风险 3	风险 1	风险 2	风险 3
过渡	1 073	339	574	1 073	339	574
RR	0.472	−0.092	−0.600	0.504	−0.185	−0.562
	(0.601)	(0.976)	(0.725)	(0.614)	(1.025)	(0.744)
DR	−0.575	−0.959	1.122	−0.806	−1.051	1.078
	(0.762)	(1.247)	(0.901)	(0.781)	(1.295)	(0.921)
UI	−1.424	−1.047	−0.966	−1.544	−1.092	−0.963
	(0.249)	(0.524)	(0.449)	(0.258)	(0.544)	(0.456)
RRUI	0.966	−0.669	−0.432	1.057	−0.742	−0.482
	(0.612)	(1.192)	(1.014)	(0.627)	(1.23)	(1.033)
DRUI	−0.198	1.987	2.102	−0.012	2.18	2.158
	(1.019)	(1.727)	(1.303)	(1.041)	(1.788)	(1.323)
LNWANG	0.351	−0.257	0.003	0.373	−0.321	−0.007
	(0.116)	(0.179)	(0.145)	(0.118)	(0.191)	(0.147)
TENURE	0	0.005	−0.047	0.000 6	0.007	−0.047
	(0.006)	(0.013)	(0.012)	(0.007)	(0.014)	(0.012)
−ln L		5 693.63			5 687.64	

19.5.1　*竞争风险框架下的估计*

若将含有异质性的指数模型与没有异质性的指数模型两两对比，则显示由于引入不可观测异质性而导致了对数似然的改进。这种结果类似于 18.8 节所报告的形式。不过，与指数模型的情形相比，含有异质性的威布尔模型有较高的对数似然，即−5 666，而前者为−5 693。含有逆高斯异质性的威布尔模型有最高的对数似然−5 543，从而看起来似乎是四个模型中最佳的。这一点不应该被解释成，对于推断来说它是一个令人满意的模型，因为该问题仍未解决。因此，我们将讨论表 19.3 的结果。

表 19.3 失业持续期限:含有 IG 脆弱性与没有 IG 脆弱性的威布尔模型竞争风险估计值与独立风险估计值

	没有异质性			IG 异质性			考克斯模型		
风险系数	风险 1	风险 2	风险 3	风险 1	风险 2	风险 3	风险 1	风险 2	风险 3
过渡	1 073	339	574	1 073	339	574	1 073	339	574
RR	0.448	−0.085	−0.694	0.736	−0.379	−0.432	0.522	−0.071	−0.469
	(0.638)	(0.992)	(0.763)	(0.906)	(1.452)	(1.111)	(−0.752)	(0.951)	(0.715)
DR	−0.472	−0.938	1.361	−1.072	−1.689	1.167	−0.571	−1.023	0.875
	(0.809)	(1.279)	(0.969)	(1.149)	(1.78)	(1.378)	(0.721)	(1.193)	(0.878)
UI	−1.496	−1.109	−1.097	−2.574	−2.063	−1.761	−1.317	−0.906	−0.905
	(0.264)	(0.527)	(0.46)	(0.384)	(0.747)	(0.623)	(0.237)	(0.510)	(0.444)
RRUI	1.015	−0.616	−0.305	1.734	−0.301	−0.515	0.882	−0.781	0.539
	(0.646)	(1.204)	(1.047)	(0.933)	(1.702)	(1.418)	(0.582)	(1.166)	(1.002)
DRUI	−0.299	1.973	1.991	−0.06	3.263	3.669	−0.095	2.031	2.293
	(1.065)	(1.757)	(1.37)	(1.538)	(2.47)	(1.935)	(0.997)	(1.671)	(1.274)
LNWANG	0.366	−0.243	0.043	0.576	−0.494	−0.006	0.335	−0.280	−0.014
	(0.122)	(0.183)	(0.153)	(0.177)	(0.261)	(0.216)	(0.110)	(0.173)	(0.141)
TENURE	−0.001	0.005	−0.049	0.0009	0.017	−0.067	0.000	0.005	−0.046
	(0.007)	(0.013)	(0.013)	(0.01)	(0.019)	(0.017)	(0.006)	(0.012)	(0.011)
α	1.29	1.08	1.17	1.75	1.65	1.79	—	—	—
	(0.022)	(0.033)	(0.028)	(0.04)	(0.06)	(0.048)	—	—	—
$-\ln L$		5 666.13			5 543.33				

威布尔模型引进不可观测异质性，导致了全部三个风险函数中风险函数斜率系数估计值的增大。就风险 1 而言，这种系数从 1.29 增大到 1.75，而对风险 2 来说，则从 1.08 增大到 1.65。也就是说，引入不可观测异质性引起了持续期限相依性大幅减小，或者骤然大幅增加失业的风险。这些变化沿着 18.5 节分析预测的线索展开。在威布尔模型中，加入不可观测异质性对失业保险（UI）系数的影响同样是相当大的，就绝对数值大小而言，实质上变得较大。RR、DR、RRUI 以及 DRUI 的系数仍然不能精确地得以确定。第一个风险函数中的 LNWANG 系数是显著的且正的，而第二个风险函数的系数则不是。也就是说，LNWANG 系数的增大促使了那些寻找全日制就业的脱离失业的过渡，却忽略了对转向兼职就业的那些人的影响。这个例子说明，竞争风险框架如何区分各种不同风险函数中变量的不同作用。

同理，考察 19.2 节给出的竞争风险模型的考克斯模型的设定。在这种设定中，不可观测异质性被忽略掉，并且基准风险不是以参数形式设定的，却如同 17.8.3 节所解释的那样可被估计出来。与表 19.2 中的那些指数模型相比，表 19.3 中的最后三列给出了点估计值，但其标准误差却很大，这是因为考克斯设定与指数模型设定相比更缺乏约束。失业保险的估计系数更接近于指数模型的而不是威布尔- IG 模型的估计系数；后者几乎是前者的 2 倍。威布尔- IG 模型中的 LNWANG 系数也较大。不过，倘若人们忽略不可观测异质性，则不可能识别基准风险。图 19.1 与图 19.2 分别表明，对于三个指定状态来说，计算基准生存函数与累积风险函数，但是这些可被更好地解释为：反映出不可观测异质性与持续期限相关的某种未知混合。这些估计值显示，那些过渡到全日制就业的基准生存函数是最低的，且位于其他两个基准生存函数的下面，同时对于过渡到兼职就业的基准生存函数，它是最平坦的并且最高。相应地，那些过渡到全日制就业的累积风险函数则是三个当中最陡峭的。

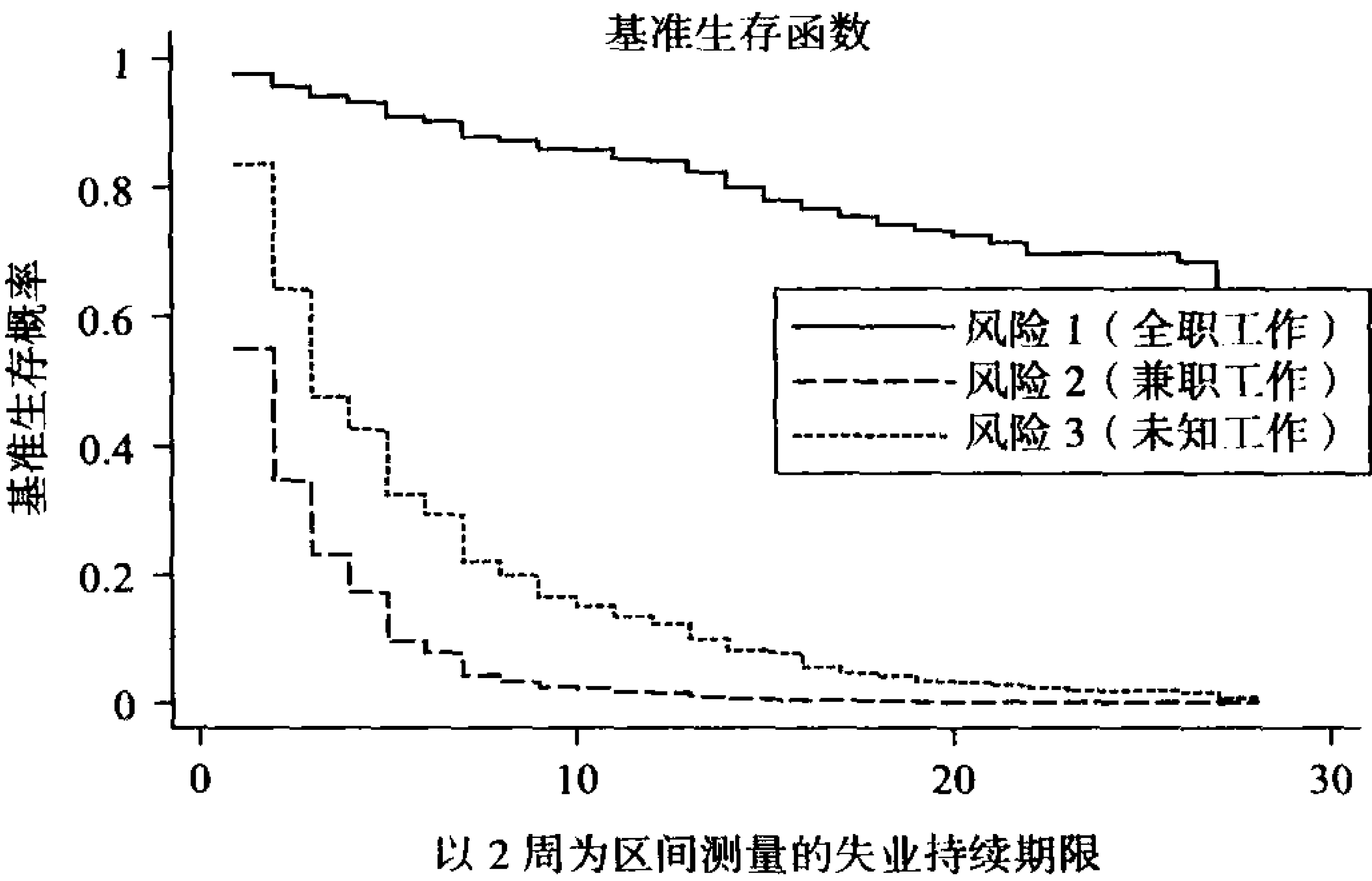


图 19.1 失业持续期限：来自考克斯竞争风险模型的估计基准生存函数。美国数据从 1986～1992 年，共计 3 343 个时期，某些是未完成的。

这里的讨论与分析仅仅是一种说明性的，从任何意义上讲不是终极性的。实际上，有好的理由表明，威布尔风险函数是一种错误设定。运用同样数据，麦考尔

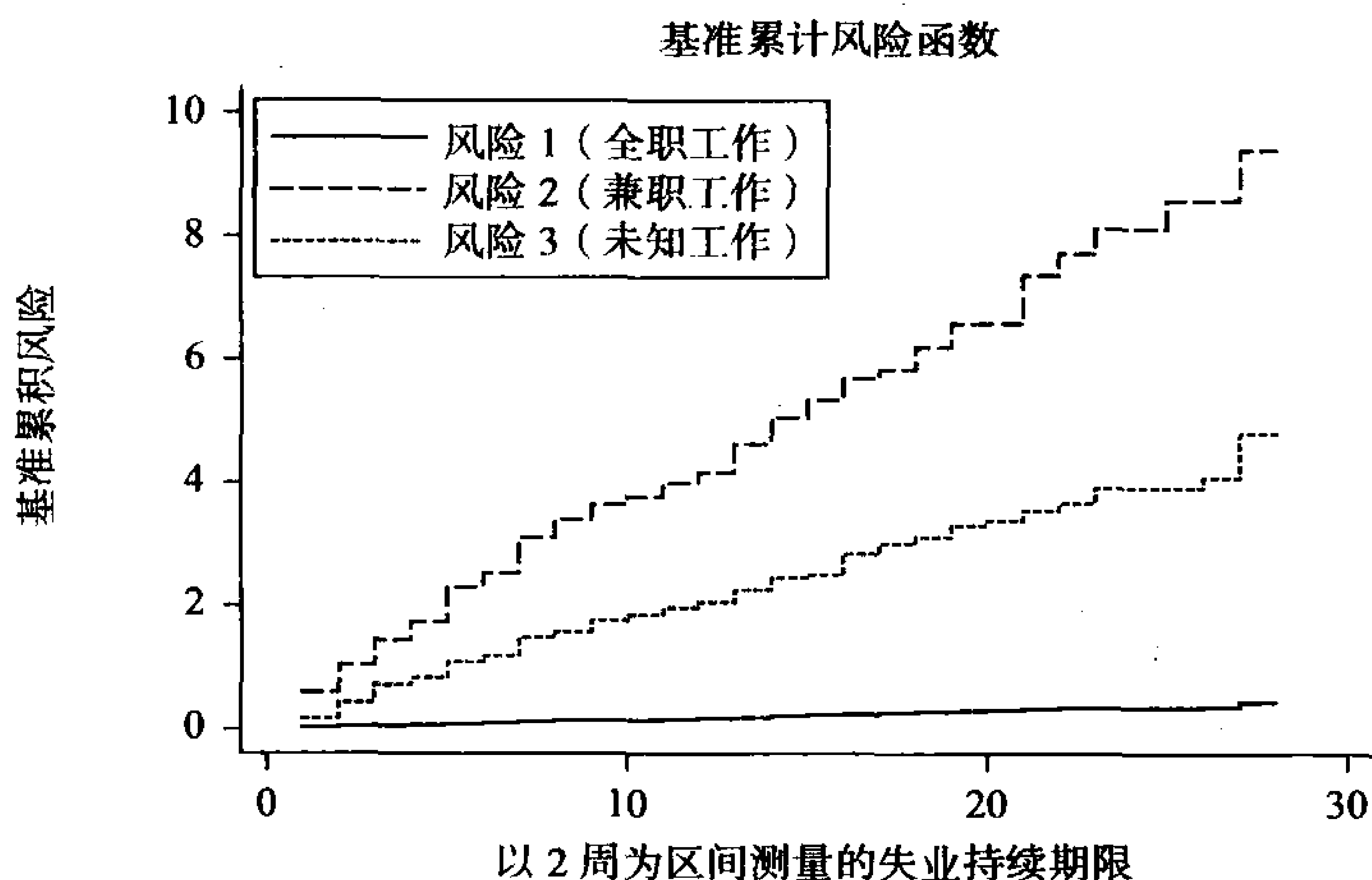


图 19.2 失业持续期限：来自考克斯竞争风险模型的估计基准生存函数。数据与图 19.1 的一样。

(McCall，1996)分析了考虑到一种更加灵活的多项式风险函数，并提供支持浴缸形状风险的证据，这意味着风险递减持续期限至低点，然后恒定不变，最终风险增大至高点持续期限。单调威布尔风险函数并没有捕获到该种可能性。其他研究者利用美国数据对失业持续期限的建模表明：当对风险函数灵活设定时，引入不可观测异质性对其结果并没有大的影响[迈耶(Meyer，1990)；哈恩和豪斯曼(Han and Hausman，1990)]。我们在这里没有看到如下事实：该事实应鼓励运用更灵活的设定，诸如 17.10 节所分析的情况。

19.6 应用研究

在对多变量生存模型建模时，一种实用方法是，在开始联立估计之前以边缘模型开始。这种策略对于评定最初设定的统计适宜性方面是有益的。

在开始研究时，多变量生存模型与风险模型的统计运算，在绝大多数情况下都需要研究者自己编程，通过使用支持软件诸如针对用户所定义的函数极大化或极小化的最优程序，能够很容易部分完成任务，这里，用户借助于许多程序与编程平台使用函数和编程语言。

含有独立风险的 CRM 简化了一系列生存模型的估计，其原因在于实际运用的信息已由 17.12 节给出。一般的多变量 CRM 程序，很难在商业软件包中找到。可是，有支持含有特殊相关结构的某些多变量生存模型。例如，STATA 支持共享脆弱性模型的计算。共享脆弱性模型(**shared frailty model**)是一种随机效应模型，对个体组或时期组来说，不可观测异质性的一些元素是共同的或部分共同分享的，并对不同组来说是随机分布的。

如果主要关注内容在于对持续期限之间的相关结构进行建模，那么联接方法相对于二变量情况的极大模拟似然法来说，潜在地更为吸引人，因为联接方法不需要数值积分。对于维数高于二维的情形，如同多重时期模型情况一样，可运用联接

方法，但已出版文献中仅有相对很少的例子。边缘模型能利用标准的一元生存模型来拟合与检验，同时相关参数运用序贯二阶段方法加以估计。即使所有参数都可联立估计，估计边缘模型也为迭代计算提供了一系列的初始值。我们没有发现支持这些模型估计的统计软件包。

19.7 文献注释

19.2 哈恩和豪斯曼(Han and Hausman, 1990)给出 CRM 的一个例子，其中，设定被推广到考虑不可观测异质性。在具有特定状态随机效应的 CRM 框架下，麦考尔(McCall, 1996)分析了某些政策变量对被保险失业者寻找兼职工作行为的影响，他运用了含有相关风险的 CRM 模型。巴特勒、安德森和伯克豪泽(Butler, Anderson, and Burkhauser, 1989)运用含有相关风险的 CRM，对接受工作风险与临终风险加以建模。

19.3 斯克拉在 1959 年以法文形式发表了关于联接的原创性文章，后来斯克拉(Sklar, 1973)的论文是以英文发表的一篇优秀论文。拉杜洛维奇和韦格坎普(Radulovié and Wegkamp)(出版年代不详)提出了斯克拉定理的一种证明。弗里斯和瓦尔德斯(Frees and Valdez, 1998)对联接文献给出一个非常有益的指导性概览，并做了文献评注。

19.4 米利和帕德尼(Mealli and Pudney, 1996)、弗林和赫克曼(Flinn and Heckman, 1982)均对多重时期进行了探索。米利和帕德尼(Mealli and Pudney, 1996)运用基于模拟的估计方法，对有权享受养老金工作、无权享受养老金工作以及其他劳动力市场状态之间的过渡进行了分析。

习 题

19-1 [改编自萨普兰(Sapra,2000;2001)。]这个问题涉及阐明 19.2 节提及的竞争风险结果的考克斯—齐亚齐斯(Cox - Tsiatsis)非识别的一个例子。考虑下述相关竞争风险模型，其中，我们观测到 $T=\min(T_1,T_2)$ 与 δ，当 $T=T_1$ 时，$\delta=1$；而当 $T=T_2$ 时，$\delta=2$。这里，T_1 与 T_2 分别表示风险 1 与风险 2 的潜在持续期限。假定二变量联合生存函数是 $S(t_1,t_2)=\exp[-(\lambda_1 t_1+\lambda_2 t_2)^\alpha]$，$0<\alpha\leqslant 1$，$\lambda_1,\lambda_2>0$。建立一个独立 CRM，该 CRM 等价于特定的相关竞争风险模型。

19-2 对于上面问题中的特定模型，若不仅 T 是可观测的，而且 δ 也是可观测的，请用风险率与综合风险率写出每一种模型的对数似然函数。考察参数信息矩阵，并证明所有参数都是局部可识别的，因为信息矩阵是非奇异的。

19-3 考察两个平行持续期限，比如说失业持续期限 T_1，与没有个人健康保险时期的持续期限 T_2，假定以不可观测异质性为条件的这两个持续期限是独立的，并且分别是均值为 $\beta_0+\beta_1 x$ 与 $\gamma_0+\gamma_1 x$ 的指数分布。假定这两个持续期限模型的乘法不可观测异质性项是 ν_1 与 ν_2，满足 $\mathrm{E}[\nu_1]=\mathrm{E}[\nu_2]=1$。

(a) 对于你选择的参数值，请写出一个算法生成(ν_1,ν_2)的相关实现值，使得不

以(ν_1,ν_2)为条件却以 x 为条件的两个持续期限将是相关的。你可随意依据数学方法或其他立式来对(ν_1,ν_2)联合分布做出引人注目的分布假设。请解释你是如何控制两个持续期限间相关的范围的。

(b) 运用 19.3.2 节给出的求二变量联合分布的方法,推导两个持续期限的联合分布。

(c) 描述你如何将(b)部分的分析加以推广,以便考虑右删失持续期限的存在。

19-4 使用与第 18 章的麦考尔数据集相同的子样本,运用含有两个状态的失业与就业的两状态模型进行估计(也就是说,忽略作为两个可供选择的指定状态的兼职就业与全日制就业之间的差异)。

(a) 用单方程威布尔模型进行拟合,并将其结果与含有威布尔设定的独立 CRM 的那些结果加以比较。

(b) 评估运用 CRM 设定而引起的拟合优度的改进。

(c) 计算并比较源于单方程与 CRM 模型的在解释变量样本平均值处计算的失业风险的拟合值。

20 计数数据模型

20.1 引 论

在许多经济背景下，关注的因变量或响应变量是非负的整数或者计数数据，我们想要利用回归元[1]（**regressor**）解释和分析它们。与经典回归模型不同，响应变量是离散的，其分布仅仅在非负的整数值上具有概率质量。本书前面曾讨论的几种模型，诸如二值结果模型与持续期限模型，都和计数数据回归模型密切相关。计数回归模型如同其他受限因变量或离散因变量模型譬如 logit 与 probit 模型一样，都是非线性的，具有与离散性及非线性密切关联的许多性质和特殊特性。

本章以独立的横截面观测值的样本数据开始，考察微观经济计量学的一些例子。生育力研究经常在对母亲年龄设定区间上对分娩孩子的数量进行建模，关注利用譬如母亲学历、年龄以及家庭收入来分析其变异情况[温克尔曼（Winkelman，1995）]。在一些家庭决策的模型里，孩子数量可作为解释变量出现，该变量是内生的。事故分析研究则通过航空公司在某时期发生的事故数量来测算航空公司安全性，并以此建模，试图确定它和航空公司赢利性以及其他航空公司财务状况测量值的关系[罗斯（Rose，1990）]。娱乐需求研究，通过对去娱乐场所出行次数进行建模，试图求出自然资源，例如国家森林的价值[古尔穆和特里维迪（Gurmu and Trivedi，1996）]。健康需求研究对个人消费健康服务的次数数据，诸如医生出诊或去年在医院住院的天数[卡梅伦等人（Cameron et al.，1988）]。如果我们想要对这种变量与一些因素，例如健康状况与健康保险之间关系进行分析，那么计数回归也是与之有关的。

20.2～20.5 节阐述主要建模方法。20.2 节详述泊松回归模型。20.3 节给出源于著名 RHIE 数据的一个应用。泊松回归模型经常显得约束性太强了，而 20.4 节阐述其他一些更普遍使用的完全参数计数模型。本节还阐述较少使用的可供选择计数模型，比如离散选择模型。20.5 节详细讨论对条件均值与条件方差进行建模的部分参数方法[2]（**partially parametric approach**）。20.6 节提供多变量计数模

[1] 又称为回归量。——译者注

[2] 又称为偏参数方法。——译者注

型以及含有内生回归元的模型。20.7 节通过利用 RHIE 数据阐述各种不同模型。随后，讨论一些实际问题。处于教学上的考虑，以某种详细方式介绍横截面数据的回归模型。许多其他优于泊松模型的一些模型，因为章节空间所限只好简略介绍。对于更完整研究，参见卡梅伦和特里维迪（Cameron and Trivedi，1998）以及文献注释。

20.2 基本计数数据回归

在一些情况下，诸如分娩生孩子数量，计数是最终关注的变量。而在另外一些情况下，比如医疗需求以及研究和发展支出的结果，最终关注的变量是连续的，这些经常花费或收入以美元来测算，但是最合适的可利用数据反而是计数数据。在许多情况下，样本集中在几个小的离散值（**few small discrete values**）上，比如说 0、1 和 2。表 20.1 列出几种发表的经济计量模型观测到的零计数比例阐明这点。这些比例在某些情况下可高达 90%。而且，数据都向右偏斜（**skewed to the right**）。最后，数据显示出内生的异方差性（**heteroskedastic**），其方差随均值变化而增大。

表 20.1　部分节选研究的零计数比例

研究	变量	样本量	零比例
卡梅伦等（1989）	就医次数	5 190	0.798
波尔迈耶和乌尔里克（1995）	专家出诊	5 096	0.678
格鲁特多斯特（1995）	处方药	5 743	0.224
德布和特里维迪（1997）	住院天数	4 406	0.806
格目和特里维迪（1996）	娱乐旅游	659	0.632
盖尔等人（1997）	住院治疗	30 590	0.899
格林（1997）	重要损毁报告	1 319	0.803

20.2.1 *泊松回归*

泊松回归是计数数据分析的起点，尽管它经常显得不合适。在 20.2.1～20.2.3 节，我们阐述泊松回归模型，这已在前面 5.2 节做了介绍，并且通过极大似然法加以估计，对估计系数给出解释，而且可推广到截尾与删失数据上。在 20.2.3 节，我们还阐述基于含有正确设定条件均值却可能错误设定条件方差的泊松分布的伪 MLE。泊松模型的局限性，即著名的等分散性质将在 20.2.4 节加以阐述。

存在一种限定条件。在一些情况下，样本中零的较高比例与非常大的计数值共存，这产生了建模上富于挑战性的困难。表 20.2 列出对专利计数与研发（R&D）支出之间的关系进行研究的 5 个信息来阐明这种特性。可以发现，最大的计数观测值会如此密切地与样本均值有关。建模上的挑战是选择一种函数形式，该函数形式能够适当地捕获到大的均值与高的零比例。在许多其他例子中，比如分娩生孩子数量，所有数据本质上被限制在单个数字上，而且事件的均值数是非常小的。

这些特性激发了对计数回归的特殊方法及模型的应用。目前,存在两种研究方法。

表 20.2 最近专利研发(R&G)研究中所用的数据集概括

研究	样本量	均值	标准差	最大专利数	零比例
钦切拉 (1997)	181	60.8	721.6	925	<0.19
克雷蓬和迪盖特 (1997b)	698	11.6	na[a]	na	0.441
克雷蓬和迪盖特 (1997a)	451	2.73	11.45	na	0.729
豪斯曼等人 (1984)	346	32.1	66.36	515	0.220
王等人 (1998)	70	23.46	39.10	173	0.186

[a] na:不可用。

第一种方法是完全参数(**fully parametric**)的方法,即完全设定数据的分布,完全将 y 限制在取非负整数值上。这种方法在早期应用中得到采用,大部分是在生物统计学里,计数回归可以被看成是对文献中大量关于独立同分布的计数分布进行扩展和推广。豪斯曼等人(Hausman et al., 1984)在其有影响的经济计量研究中也采用了这种方法。

第二种方法是均值方差方法(**mean-variance approach**),即设定条件均值是非负的,并设定条件方差是条件均值的函数。这种方法充分地对非负性与异方差性进行建模,但没有讨论数据的离散性。这个方法在没有受限且仅为计数数据框架下,由内尔德和韦德伯恩(Nelder and Wedderburn, 1972)引入,后来导致了统计学被广泛运用的广义线性建模方法[麦卡拉和内尔德(McCullagh and Nelder, 1989)]。在经济计量学中,该方法是由古里耶克斯、蒙福特和特罗格恩(Gourieroux, Monfort, and Trognon, 1984a, b)引入的,最好是将它看成对广义矩方法的专门研究。

20.2.2 *泊松 MLE 与 QMLE*

第 5 章已经引入并讨论的泊松 MLE 与拟 MLE(QMLE)可作为 m 估计的一个例子。这里我们给出更完整的研究。

关于计数的一个自然而然的随机模型是,关注事件发生的泊松点过程。这蕴含着事件发生数的泊松分布(**Poisson distribution**)具有密度,或更正式地讲,概率质量函数为:

$$\Pr[Y=y]=\frac{e^{-\mu}\mu^{y}}{y!},\quad y=0,1,2,\cdots, \tag{20.1}$$

其中,μ 表示强度或速率参数。我们将该分布称为 $\mathcal{P}[\mu]$。它的前二阶矩是:

$$\begin{aligned}\mathrm{E}[Y]&=\mu\\ \mathrm{V}[Y]&=\mu\end{aligned} \tag{20.2}$$

这表明泊松分布的著名等分散(**equidispersion**)(均值和方差相等)性质。

通过引入观测值下标 i,即附于 y 和 μ 上,iid 框架可被推广到回归情况。通过对均值参数 μ 与协变量(回归元)$\mathbf{x}$ 之间关系进行参数化,由泊松分布可推导出泊松回归模型(**Poisson regression model**)。标准假设是使用指数均值参数化:

$$\mu_i=\exp(\mathbf{x}_i'\boldsymbol{\beta}),\quad i=1,\cdots,N \tag{20.3}$$

由假设知，存在 K 个线性独立的协变量，通常包括常值。因为 $\mathrm{V}[y_i|\mathbf{x}_i]=\exp(\mathbf{x}_i'\boldsymbol{\beta})$，由式(20.2)与式(20.3)知，泊松回归具有内在的异方差性。

给定式(20.1)与式(20.3)，以及观测值$(y_i|\mathbf{x}_i)$是独立的假设，最自然的估计量是极大似然估计量。其对数似然函数是：

$$\ln \mathrm{L}(\boldsymbol{\beta}) = \sum_{i=1}^{N}\{y_i\mathbf{x}_i'\boldsymbol{\beta} - \exp(\mathbf{x}_i'\boldsymbol{\beta}) - \ln y_i!\} \tag{20.4}$$

泊松 MLE(**Poisson MLE**)记为 $\hat{\boldsymbol{\beta}}_{\mathrm{P}}$，是对应于极大似然的一阶条件的 K 个非线性方程：

$$\sum_{i=1}^{N}(y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{x}_i = \mathbf{0} \tag{20.5}$$

如果 $\mathbf{x}_i$ 包括常数项，那么由式(20.5)知，残差 $y_i-\exp(\mathbf{x}_i'\boldsymbol{\beta})$和为 1。其对数似然函数是全局凹的；因此，要想求解这些方程，通过高斯—牛顿或牛顿—拉夫森迭代算法可以得到唯一的参数估计值。

在经济计量学文献中，伪 MLE(**pseudo-ML**, PML)或准 ML(**quasi-ML**, QML)估计意指，在对错误设定密度下通过 ML 来进行估计[古里耶克斯等人(Gourieroux et al.,1984a)]。PML 与 QML 术语经常可以交换使用。在数据生成过程的假设下，可获得该估计量分布，而关于数据生成过程的假设比导致特定似然函数的假设要弱一些；参见 5.7 节。在统计文献中，QML 常常意指非线性广义最小二乘法。对于泊松回归来说，QML 在后者的意义下等价于标准极大似然法。

由式(20.5)，泊松 PML 估计量 $\hat{\boldsymbol{\beta}}_{\mathrm{P}}$ 具有一阶条件$\sum_{i=1}^{N}(y_i-\exp(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{x}_i=\mathbf{0}$。正如已经注意到的，当 $\mathrm{E}[y_i|\mathbf{x}_i]=\exp(\mathbf{x}_i'\boldsymbol{\beta})$时，左边之和等于期望 0。因此，在对条件均值正确设定的较弱假设下，泊松 PML 是一致的；也就是说，数据不需要服从泊松分布。利用 5.2.3 节给出的结果，方差矩阵就是三明治形式的，满足：

$$\mathrm{V}_{\mathrm{PML}}[\hat{\boldsymbol{\beta}}_{\mathrm{P}}] = \left(\sum_{i=1}^{N}\mu_i\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^{N}\omega_i\mathbf{x}_i\mathbf{x}_i'\right)\left(\sum_{i=1}^{N}\mu_i\mathbf{x}_i\mathbf{x}_i'\right)^{-1} \tag{20.6}$$

并且，$\omega_i=\mathrm{V}[y_i|\mathbf{x}_i]$表示 y_i 的条件方差。

由标准的 ML 理论，如果较强的假设使得泊松回归在参数形式上得以正确设定，所以 $\omega_i=\mu_i$，那么估计量 $\hat{\boldsymbol{\beta}}_{\mathrm{P}}$ 关于 $\boldsymbol{\beta}$ 是一致的，而且是渐近正态的，具有样本协方差矩阵：

$$\mathrm{V}[\hat{\boldsymbol{\beta}}_{\mathrm{P}}] = \left(\sum_{i=1}^{N}\mu_i\mathbf{x}_i\mathbf{x}_i'\right)^{-1} \tag{20.7}$$

在此情况下，μ_i 具有指数形式(20.3)。

泊松 ML 估计量与 PML 估计量是一样的，却具有不同方差。20.5.1 节将阐述对更稳健估计(20.6)的实证例子。

20.2.3 *解释回归系数*

对于满足 $\mathrm{E}[y|\mathbf{x}]=\mathbf{x}'\boldsymbol{\beta}$的线性模型来说，系数 $\boldsymbol{\beta}$ 已经被解释成为回归元变化一个单位对条件均值的效应。对于非线性模型而言，需要对此加以修改。参见 5.2.4 节给出的一般性讨论。对具有指数条件均值的任何模型来说，对其微分得到：

$$\frac{\partial \mathrm{E}[y|\mathbf{x}]}{\partial x_j}=\beta_j \exp(\mathbf{x}'\boldsymbol{\beta}) \tag{20.8}$$

其中，纯量 x_j 表示第 j 个回归元。例如，若 $\hat{\beta}_j=0.25$ 且 $\exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}})=3$，第 j 个回归元变化一个单位时，则引起 y 的期望值增加 0.75 单位。这种偏响应依赖于 $\exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}})$，它对于不同的个体而言预期是变化的。容易理解，$\beta_j$ 测算由 x_j 变化一个单位时引起 $\mathrm{E}[y|x]$ 的相对变化。如果 x_j 在对数标度上进行测算，那么 β_j 就是弹性的。

为了报告单个响应值，一个好的备选者是平均响应估计值，$N^{-1}\sum_i \partial \mathrm{E}[y_i|\mathbf{x}_i]/\partial x_{ij}=\hat{\beta}_j\times N^{-1}\sum_i \exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}})$。对于含有截距的泊松回归模型来说，可以证明，这可简化成 $\hat{\beta}_j\bar{y}$ 。

式(20.8)的另一个结果就是，比如说，如果 β_j 是 β_k 的 2 倍，那么第 j 个回归元变化一个单位而引起的效应，就是第 k 个回归元变化一个单位而引起效应的 2 倍。

20.2.4 *过度分散*

就计数数据而言，泊松回归模型通常约束性太强，这导致了由 20.3 节与 20.4 节所阐述的一些其他可供选择的模型。基本问题是，其分布要用纯量单参数(μ)来参数化，所以 y 的所有矩都是 μ 的函数。与之相比，正态分布具有位置(μ)与标度(σ^2)各自分开的参数。由于同样的原因，对计数数据而言，单参数的指数约束性太强，而更一般的两个参数分布，诸如威布尔分布就表现得优越一些。注意到，对于二值数据来说，这种复杂性不会产生。于是，如果成功概率是 p，那么失败概率必是 $1-p$，显然分布是一个参数的贝努里分布。不过，对于二值数据而言，问题是如何用回归元去参数化 p。

表现这种约束性的一种方式是，在许多应用中，泊松密度预测零计数的概率相当小于在样本中所实际观测到的。这称为超额零(**excess zeros**)问题，因为数据中的零比泊松预测要更多些。

泊柏模型的第二个明显的不足之处是，对计数数据而言，其方差通常大于均值，此特性称为过度分散(**overdispersion**)。相反，泊松模型蕴含，其方差与均值是相等的[参见式(20.2)]，这个性质称为等分散性。

从性质上看，过度分散具有类似于线性回归模型中同方差性假设失败的结果。倘若条件均值得到正确设定，即式(20.3)成立，泊松 MLE 还是一致的。由于如果 $\mathrm{E}[y_i|\mathbf{x}_i]=\exp(\mathbf{x}_i'\boldsymbol{\beta})$，那么式(20.5)左边将具有零期望，所以这没有对式(20.5)的一阶条件进行检验。当设定密度处于 LEF 之中时，这种一致性更一般地用于拟 MLE。不仅泊松分布，而且正态分布都是前面 5.7.3 节曾讨论的 LEF 的成员。不过，重要的是控制过度分散。首先，在更复杂背景下，诸如含有截尾与删失情况，过度分散会导致更基本的非一致性问题。其次，甚至在最简单背景下，较大的过度分散会导致极度缩小标准误差且极大夸张 t 稳健方差估计量。再次，如果人们想要估计事件数的概率而不仅仅是条件均值，这些都依赖于额外的数据生成过程参数。

过度分散可作为更基本的错误设定存在的信号，尤其是在涉及截尾与删失的背景下，在估计时将它们忽略掉。在这种情况下，条件均值被错误设定，并且过度分散联立存在，这将导致 MLE 的无效性以及非一致性。

因此，在实施泊松回归之后，对过度分散进行统计检验是人们非常期望的。大部分含有过度分散的计数模型，把过度分散设定成为如下形式：

$$V[y_i \mid \mathbf{x}_i] = \mu_i + \alpha g(\mu_i) \tag{20.9}$$

其中，α 表示未知参数，而 $g(\cdot)$ 表示已知函数，最普遍的是 $g(\mu)=\mu^2$ 或 $g(\mu)=\mu$。假定既在零假设下又在备择假设下，均值都被正确设定。例如 $\exp(\mathbf{x}_i'\boldsymbol{\beta})$，而在零假设下 $\alpha=0$，因此，$V[y_i \mid \mathbf{x}_i]=\mu_i$。关于 H_0：$\alpha=0$ 与 H_1：$\alpha\neq 0$ 或 H_1：$\alpha>0$ 的一种简单过度分散检验统计量（**overdispersion test statistic**），建立拟合值 $\hat{\mu}_i=\exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}})$，并实施辅助 OLS 回归（不含常值的）：

$$\frac{(y_i-\hat{\mu}_i)^2-y_i}{\hat{\mu}_i}=\alpha\frac{g(\hat{\mu}_i)}{\hat{\mu}_i}+u_i \tag{20.10}$$

其中，u_i 表示误差项，能通过估计泊松模型得到计算出来。在没有过度分散的零假设下[卡梅伦和特里维迪（Cameron and Trivedi, 1990）]，即使这里可使用生成的回归元，所报告的关于 α 的 t 统计量是渐近正态的。这种检验还可用于分散不足（**underdispersion**），即 $\alpha<0$，在此情况下，条件方差小于条件均值。还可参见古尔穆和特里维迪（Gurmu and Trivedi, 1992）。

20.3 计数例子：就医次数

为了阐明理论，我们使用源于 RAND 健康保险实验的一些数据，这些数据以前曾被德布和特里维迪（Deb and Trivedi, 2002）使用。和这里所给的阐述相比，他们对模型进行了更完整的估计，并完成对数据较为深刻的分析，由 RAND 公司领导的这个实验是从 1974～1982 年，这在医疗保健研究（medical care research）中是实施最长且最大的控制性社会实验。实验的主要目的是评估病人对健康服务使用如何受到各类随机安排健康保险的影响，既包括一次一付医疗费[1]（**fee-for-service**），又包括健康维护组织（HMO）。在该实验中，数据是从 2 823 个家庭中的 8 000 个入会者那里搜集的，它们来自全国的 6 个城市。每一个家庭要在 3 年或 4 年的 14 种各种不同健康保险计划中注册一种。计划范围从自由照料到 95%共同保险低于最高美元开支（maximum dollar expenditure, MDE），同时还包括被预付小组实践中的安排。

重要的核心思想是，由于保险计划是随机安排的，并不可以由参与者自由地选择，所以我们没有面对内生处理效应问题，这是正在研究关注的中心因果参数（原因参数）。

数据是从入会者使用医疗保健服务以及对于或 3 年或 5 年入会项目的随机安排的健康状态所搜集而来的。有关该数据的更多详细内容，参见曼宁等人（Manning et al., 1987）、纽豪斯等人（Newhouse et al., 1993），以及德布和特里维迪（Deb and Trivedi, 2002）。在该研究中，所使用的样本是由仅仅参与一次一付医疗费的个体组成。

---

[1] 又称为按服务项目收费。——译者注

数据文件是由利用(效用)、开支、人口特征、健康状况以及保险状态变量组成的。开支数据已在16.6节中进行了分析。此样本共保率(**coinsurance rate**)假定四种不同的数值。然而,遵从RAND研究,我们把它处理成为一个连续变量。最终样本由20 186个观测值组成,每一个观测值代表给定年份中一个实验题目的数据。为了简单起见,这里忽略数据中出现的集群[1](**clustering**),参见24.5节。

在目前阐述中,所利用的分析测量就是与医生联系次数(MDU)。以百分比形式给出的MDU的相对频率分布已由表20.3给出。MDE表示最大美元开支(**maximum dollar expenditure**),实际中有一个医疗开支债务限制,在此限制之上,参与者将不负担成本分摊。观察发现,大致31%的观测值是0。较长的右尾与方差非常大于均值,这表明计数是(无条件)过度分散。

表20.3 就医次数:频率分布

接触	0	1	2	3	4	5	6	7	8	9	10
相对频数	31.2	18.9	13.8	9.3	6.7	4.8	3.4	2.6	2.0	1.4	1.0
接触	11	12	13	14	15	16	…	>21	Max		
相对频数	0.9	0.6	0.5	0.4	0.3	0.3		1.0	77		

这里讨论的目的是,我们考察通过泊松ML与泊松PML进行回归估计。其他的设定则稍后考虑。就一切情况而论,所包括的协变量是表20.4中的那些。

表20.4 就医次数:变量说明

变量	定 义	均值	标准差
MOU	门诊病人访问MD人数	2.861	4.505
LC	ln(共保险+1),0≤共保险≤100	1.710	1.962
IDP	若个人可减免的,则取1,否则取0	0.220	0.414
LPI	ln(max(1,年度参与激励支付))	4.709	2.697
FMDE	当IDP=1时,为0	3.153	3.641
	ln(max(1,MDE/(0.01共保))),其他		
LINC	ln(家庭收入)	8.708	1.228
LFAM	ln(家庭人口数)	1.248	0.539
AGE	年龄	25.718	16.768
FEMALE	当此人为妇女时,取1	0.517	0.500
CHILD	当年龄小于18岁时,取1	0.402	0.490
FEMCHILD	FEMCHILD * CHILD	0.194	0.395
BLACK	当户主种族是黑人	0.182	0.383
EDUCDEC	户主受教育年数	11.967	2.806
PHYSLIM	当此人受体质限制,取1	0.124	0.322
NDISESE	几种慢性病	11.244	6.742
HLTHG	若此人自测健康状况良好,取1	0.362	0.481
HLTHF	若此人自测健康状况一般,取1	0.077	0.267
HLTHP	若此人自测健康状况不好,取1	0.015	0.121
	省略分类是自测健康状况极好		

[1] 又称为聚集,详细内容参见第24章。——译者注

表 20.5 中给出对有意思的系数及其 t 比率的选取，以及对数似然与信息准则。为了节省空间，我们没有将全部内容输出重述。与保险变量（LC、JDP、LPI 以及FMDE）相联系的变量系数显然是人们关注的，因为它们反映出对价格的敏感性。此外，五个健康状况变量的系数（PHYSLIM、NDISEASE、HLTHG、HLTHF 以及 HLTHP）也是关注的内容。

表 20.5　就医次数：计数模型估计

模型	泊松		PPML	NB2-PML	
	系数	t 比率	t 比率	系数	t 比率
LC	−0.042 7	−7.030	−2.835	−0.050 4	−3.228
IDP	−0.161 3	−13.881	−5.773	−0.147 5	−4.889
LPI	0.012 8	6.999	2.912	0.015 8	3.574
FMDE	−0.020 6	−5.803	−2.319	−0.021 3	−2.351
PHYSLIM	0.268 4	21.711	8.240	0.275 1	8.068
NDISEASE	0.023 1	38.124	13.487	0.025 9	15.324
HLTHG	0.039 4	4.109	1.699	0.006 5	0.275
HLTHF	0.253 1	15.613	5.894	0.236 8	5.425
HLTHP	0.521 6	19.150	6.966	0.425 6	6.205
α	—	—	—	1.182 2	8.926
−ln L	60087			42777	

考察共保率的系数 LC，这里用对数标度进行测算。该变量是主要关注的内容，因为它提供了有关价格效应的信息。共保率越高，由病人分摊的成本就越大，从而平均就诊次数就越少。源自泊松回归所估计的系数（参见表 20.5 第 1 列）如同由标准理论所预测的，是负的（−0.042），其 t 比率为 2.835，表明价格效应显著为负的。就医次数对 LC 的弹性是−0.042。不过，由于共保率仅仅取几个少数值且没有连续变化，所以在解释这个值时应该运用保健。受限于这个限定条件，可将系数解释成为弹性。类似地，关于收入对数（LINC）是 0.174，表明收入增加会引起平均就诊次数提高。

泊松回归拟合数据程度果真会好吗？一种简单判断此问题的方法是，对于各种不同的就医次数来说，比较真实的频数与拟合的频数。表 20.6 提供了直到 9 次出诊的比较情况，而省略总体解释小于 10%出诊时的较大频数。为了计算拟合值 $\Pr[y_i|\mathbf{x}_i'\hat{\beta}]$，对于 $y_i=0,1,\cdots,9$，将 $\hat{\mu}_i$ 代入式（20.1），然后对观测值取平均。可以发现，泊松回归严重低估了零次出诊比例，而过高估计了出诊次数直到 7 次的正比例。因而，我们得出结论，泊松回归是有缺陷的。可以证明，出现该种拟合不足模式与忽略数据的过度分散有关[卡梅伦和特里维迪（Cameron and Trivedi, 1998，第 4 章）]。

表 20.6　就医次数：观测到的频率与拟合频率

接触频数	0	1	2	3	4	5	6	7	8	9
相对频数	31.2	18.9	13.8	9.3	6.7	4.8	3.4	2.6	2.0	1.4
泊松拟合	10.6	19.2	20.9	17.6	12.6	7.99	4.69	2.64	1.46	0.8
NB2 拟合	30.9	19.6	13.6	9.76	6.97	5.07	3.70	2.72	2.0	1.47

在忽略过度分散情况下，可以预见，泊松 MLE 的 t 比率将会夸大。比较表 20.5 第 3 列(PPML)的稳健 t 比率，可以证明，实际上确是如此。例如，稳健性引起 LC 的 t 比率从 -7.03 下降到 -2.83。表 20.5 与表 20.6 包括了将要在 20.7 节讨论的 NB2 模型的一些结果。对于这些数据，NB2 模型是一个好的参数模型。

20.4 参数计数回归模型

泊松回归经常表现出约束性太强。本节，我们将描述一些更灵活的可供选择的其他参数形式。

第一，计数数据中的过度分散归因于不可观测异质性。在这种情况下，计数可被看成由泊松过程生成(在此情况下，事件是序列独立的)，可是研究者没能正确设定此过程的速率参数。相反，速率参数本身就是一个随机变量。20.4.1 节与 20.4.2 节将阐述的混合方法导致了广泛运用负二项式模型。

第二，过度分散以及在一些情况下产生的分散不足，是因为生成第一个事件的过程不同于决定稍后事件的过程。例如，最初医生出诊仅仅是病人选择，而以后出诊则是由医生来决定。这就阐述 20.4.5 节所述的修正计数模型。

第三，计数数据中的过度分散可归因于对事件独立性假设的失败，它隐含于泊松过程之中。例如，人们能够假定相依性，因而一名医生出诊会使医生后来更可能出诊。(这种方法没有广泛用于计数数据分析之中。在持续期限分析里，这称为真实状态相依性。)对不可观测异质性或者相依性的特殊假设再次导致负二项式情况；参见温克尔曼(Winkelmann, 1995)。20.4.6 节将进一步阐述对 $\Pr[y=j \mid y \geq j-1]$ 建模的一种离散选择模型。

第四，人们参考对单变量 iid 计数分布的扩展与丰富文献，譬如对数序列与超几何分布[约翰逊、科茨和肯普(Johnson, Kotz and Kemp, 1992)]。通过设置一个或更多个分布参数成为回归元的设定函数来发展新的回归模型。这里将不表述这类模型。此类方法比前三种方法更缺少动机，而且得到的模型不是非常好。

尽管强调过度分散，但也会产生分散不足。例如，在计数结果可能是 0 或 1 的样本中，具有非常小的 2s 数，因此接近于二项式模型，这将表现出分散不足。卡茨分布族(Katz family of distributions)的一些成员，或者建立在级数展开方法之上的其他分布，诸如由卡梅伦和约翰逊(Cameron and Johansson, 1997)所发展起来的那些分布，都可以使用；可参见卡梅伦和特里维迪(Cameron and Trivedi, 1998，第 12 章)。

20.4.1 *负二项式模型*

负二项式模型能以许多不同方式来获得，它是连续混合模型的一个特例。下面利用混合分布的推导是最古老且具有广泛影响力的。

假定随机计数 y 的分布是泊松分布，以参数 λ 为条件，因此，$f(y \mid \lambda)=\exp(-\lambda)\lambda^{y}/y!$。现在假定参数 λ 是随机的，而不是回归 $\mathbf{x}$ 的完全确定性函数。特别地，设 $\lambda=\mu\nu$，其中，μ 表示 $\mathbf{x}$ 的确定性函数，例如 $\exp(\mathbf{x}'\boldsymbol{\beta})$，而 $\nu>0$ 是 iid 的，其

密度为 $g(\nu|\alpha)$。这是不可观测异质性的一个例子，因为各种不同观测值可具有不同的 λ（异质性），这种差异部分归因于随机（不可观测的）成分 ν。注意到，当 $\mathrm{E}[\nu]=1$ 时，$\mathrm{E}[\lambda|\mu]=\mu$，因此对斜率参数的解释如同泊松模型的一样。

y 的边际密度不是以随机参数 ν 为条件的，而是以确定性参数 μ 与 α 为条件的，它可通过积分去掉 ν 得到。从而，得到：

$$h(y|\mu,\alpha)=\int f(y|\mu,\nu)g(\nu|\alpha)d\nu \tag{20.11}$$

其中，$g(\nu|\alpha)$ 称为混合分布，而 α 表示该混合分布的未知参数。此积分定义出一种"平均"分布。对于 $f(\cdot)$ 与 $g(\cdot)$ 的某种特殊来说，积分将具有显性解或闭形式解。

当 $f(y|\lambda)$ 表示泊松密度，并且 $g(\nu)=\nu^{\delta-1}e^{-\nu\delta}\delta^{\delta}/\Gamma(\delta)$，$\nu,\delta>0$ 表示伽玛密度时，满足 $\mathrm{E}[\nu]=1$ 且 $\mathrm{V}[\nu]=1/\delta$，就得到如下作为混合密度的负二项式（**negative binomial**）：

$$\begin{aligned}h[y|\mu,\delta]&=\int_0^{\infty}\frac{e^{-\mu\nu}(\mu\nu)^y}{y!}\frac{\nu^{\delta-1}e^{-\nu\delta}\delta^{\delta}}{\Gamma(\delta)}d\nu \qquad (20.12)\\&=\int_0^{\infty}\frac{e^{-(\mu+\delta)\nu}\mu^y}{y!}\frac{\nu^{y+\delta-1}\delta^{\delta}}{\Gamma(\delta)}d\nu\\&=\frac{\mu^y\delta^{\delta}}{\Gamma(\delta)y!}\int_0^{\infty}e^{-(\mu+\delta)\nu}\nu^{y+\delta-1}d\nu\\&=\frac{\mu^y\delta^{\delta}\Gamma(y+\delta)}{\Gamma(\delta)y!(\mu+\delta)^{y+\delta}}\\&=\frac{\Gamma(\alpha^{-1}+y)}{\Gamma(\alpha^{-1})\Gamma(y+1)}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\mu+\alpha^{-1}}\right)^{y}\end{aligned}$$

其中，$\alpha=1/\delta$，$\Gamma(\cdot)$ 表示伽玛积分，即规定为整数自变量的阶乘，而第 4 行则经过某种代数运算以及利用了伽玛函数的定义之后而得到。负二项式的一些特殊情况包括，泊松分布（$\alpha=0$）、从 δ 到 α 的重新参数化优势以及几何分布（$\alpha=1$）。

如同许多混合分布情况一样，负二项式还有独立推导；参见卡梅伦和特里维迪（Cameron and Trivedi, 1998，第 4 章）。它可通过许多不同方式得到，而且人们并不总是将它看成一种混合分布。

对作为泊松伽玛混合（**Poisson-gamma mixture**）的负二项式的代数推导，能给出贝叶斯解释。给定 α 与 13.2.4 节关于指数族的共轭先验结果，μ 的先验分布是伽玛分布。可以预计，其后验分布具有闭形式。因此，在关于 α 的非确定先验（分布）的进一步假设下，MLE 与贝叶斯后验均值是一致的。

负二项分布的前两阶矩是：

$$\begin{aligned}\mathrm{E}[y|\mu,\alpha]&=\mu \qquad (20.13)\\\mathrm{V}[y|\mu,\alpha]&=\mu(1+\alpha\mu)\end{aligned}$$

因为 $\alpha>0$ 且 $\mu>0$，故方差大于均值。实际上，很容易证明，若 $y|\lambda$ 是泊松分布，且不可观测异质性具有乘法形式 $\lambda=\mu\nu$，其中，$\mathrm{E}[\nu]=1$，则总会产生过度分散。还要注意到，该过度分散具有 20.2.4 节所讨论的式（20.9）形式。

负二项式的两个标准变形经常用于回归应用中。这两个变形均设定 $\mu_i=\exp(\mathbf{x}_i'\boldsymbol{\beta})$。最普通的变形是，设 α 是待估参数，源于式（20.13）的条件方差函数

$\mu+\alpha\mu^2$关于均值是二次的。

负二项式模型的其他变形具有线性方差函数，$V[y|\mu,\alpha]=(1+\gamma)\mu$，即通过用$\gamma/\mu$代替式(20.12)的$\alpha$得到。另一方面，通过ML可直接进行估计。有时，这个变形称为负二项式1(NB1)，使之与含有二次方差函数的变形即称为负二项式2(NB2)的模型形成对比[卡梅伦和特里维迪(Cameron and Trivedi, 1998)]。很容易从式(20.12)获得对数似然。模型的这两种变形都很容易地通过ML得到估计，例如，其详细推导由卡梅伦和特里维迪(Cameron and Trivedi, 1998)给出。在这两种变形中，由于$E[y|\mathbf{x}]=\exp(\mathbf{x}'\boldsymbol{\beta})$，所以其系数具有相同解释。如同20.7节的应用一样，NB2变形是最经常运用的。

在应用研究中，发现NB2模型非常有用。它为更好拟合计数数据的许多类型提供了必要的灵活性。之所以这样，部分因为二次方差设定在许多实证情况下是一种好的近似。NB2经常提供好的拟合，该事实的一个不幸结果是，倘若泊松假设失效，一旦忽视其他可能性，则存在越过可供选择的负二项式。应避免这种机械式方法，因为泊松模型表现不好，其原因在于条件均值函数没有设定好，可以发现，运用负二项式模型保持相同条件均值。

与泊松模型相比，负二项式模型对于分布错误设定来说更缺少稳健性。即使条件均值得到正确设定，负二项式模型的MLE也是非一致的，除NB2模型的特殊情况之外，关于$\boldsymbol{\beta}$(但不是α)的MLE还是一致的。

就计数的混合模型而言，由于泊松过程是关于计数的一个正常模型，所以对于式(20.12)中的初始密度来说，泊松密度就是一个自然选择。对于式(20.12)的混合分布$g(\nu)$来说，选择伽玛分布就更具有任意性。对它的使用会产生18.2～18.4节所讨论的一些问题。其他的可能选择包括对数正态分布与逆高斯分布。参见威尔莫特(Willmot, 1987)以及郭和特里维迪(Guo and Trivedi, 2002)。在这些情况下，边缘分布不能用闭形式表述，因为它是伽玛分布，而伽玛是泊松的共轭。当然，这不意味着，所得到的模型不能由极大似然法估计。它意味着，人们必须要使用数值方法求积分或模拟极大似然法估计模型。对当前可利用的计算能力来说，这些方法总体上是可行的。假如人们准备用第12章讨论的基于模拟的估计方法，则利用各种不同类型的混合泊松模型的范围将变得非常广泛。

20.4.2 *模拟极大似然法*

为了方便理论阐述，我们现在阐明如何通过极大模拟似然法(**maximum simulated likelihood**)估计NB2模型。读者应该认识到，由于我们已拥有那个模型的解析表达式，所以在实际应用中这是多余的。假定没有模型解析式，就只好通过模拟求解估计。

注意到，式(20.12)的$h(y|\alpha,\mu)$能由

$$\frac{1}{S}\sum_{s=1}^{S}\frac{e^{-\mu\nu_s}(\mu\nu_s)^y}{y!}$$

来逼近，其中，$\nu_s(s=1,\cdots,S)$表示从分布$g(\nu|\alpha)$得到伪随机采样，而S表示所用模拟复制次数。可直接从均值为1且方差为α的伽玛分布进行采样。人们从均匀分

布采样，然后对它运用一个变换。设 u_s 表示均匀随机变量，并设 $\nu_s = -\ln u_s/\alpha$，然后定义一个模拟元：

$$\tilde{f}(y|\nu_s,\alpha,\mu) = \frac{e^{-\mu(-\ln u_s/\alpha)}(\mu(-\ln u_s/\alpha))^y}{y!}$$

于是，MSL 估计量 $\hat{\theta}_{\mathrm{MSL}}$ 极大化：

$$Q_N(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln\left(\frac{1}{S}\sum_{s=1}^{S}\tilde{f}(y_i|x_i,u_i^s,\boldsymbol{\theta})\right) \tag{20.14}$$

其中 $\mu_i = \exp(\mathbf{x}'\boldsymbol{\beta})$，而 $\boldsymbol{\theta} = (\alpha,\boldsymbol{\beta})$。

当然，这种方法是密集计算，其他情况就简单易行。对 MSL 性质的更多讨论，将由 12.4 节给出。这里，我们提醒读者，当 $S,N\to\infty$，$S/\sqrt{N}\to 0$ 时，$\hat{\boldsymbol{\theta}}_{\mathrm{MSL}}$ 与 $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ 是渐近等价的。

20.4.3 *有限混合模型*

在上面一节里，由于混合随机变量 ν 被假定具有连续分布，所以混合模型是连续混合模型。相反，一种可供选择的方法是使用不可观测异质性的离散表示，这就产生一类被称为有限混合(**finite mixture**)的模型；参见 18.5 节。该类模型是潜类型模型(**latent class models**)的特殊子类。这种模型的一些变形或特殊情况，还被统称为离散因素模型(**discrete factor models**)。

在经验研究中，对连续混合的一种可供选择的更广泛运用是下一节将讨论的修正计数模型类型。不过，更自然的是继承前面一节，对有限混合讨论。进一步地，可将修正计数模型的子类看成是有限混合的一种特殊情况。

我们假定，y 的密度是 m 个不同密度的线性组合，其中，第 j 个密度是 $f_j(y|\boldsymbol{\theta}_j)$，$j=1,2,\cdots,m$。因而，$m$ 个成分有限混合是：

$$f(y|\boldsymbol{\theta},\boldsymbol{\pi}) = \sum_{j=1}^{m}\pi_j f_j(y|\boldsymbol{\theta}_j),\quad 0\leqslant\pi_j\leqslant 1,\quad \sum_{j=1}^{m}\pi_j = 1 \tag{20.15}$$

为了一般性，在给定公式中，假定混合成分在其参数方面各不相同。更具有约束性的公式假定，仅有某些参数在不同成分上不一样(比如截距)，并且剩余参数对于混合成分来都是共同的。也可以做出某种中间层面的一般性假设。

为了进一步考察这种方法，分析 $m=2$ 情况。假定抽样总体包括两种“类型”情况，其中，y 的结果是由分布 $f_1(y|\boldsymbol{\theta}_1)$ 与 $f_2(y|\boldsymbol{\theta}_2)$ 刻画，我们假定它们具有不同的矩。假定类型 1 子总体具有均值 $\mu(\boldsymbol{\theta}_1)$，而类型 2 子总体具有均值 $\mu(\boldsymbol{\theta}_2)$，其中，$\mu(\boldsymbol{\theta}_2)<\mu(\boldsymbol{\theta}_1)$。例如，在对医疗服务的使用研究中，类型 1 子总体对应于频繁使用者，而类型 2 则对应于相对不频繁使用者。假定总体中这两种类型的部分分别是 π_1 与 $\pi_2(=1-\pi_1)$。于是，从该总体中抽取的随机样本将包括两种类型的 π_1 与 π_2 比例，尽管人们不能观测到哪一种情况属于哪一个子总体。也就是说，“类型”是潜类型(**latent classes**)。

运用这种模型的研究者是要估计未知参数 $\boldsymbol{\theta}_j$，$j=1,2,\cdots,m$。很容易发展基于式(20.15)的回归模型。例如，使用 NB2 模型，$f_j(y|\boldsymbol{\theta}_j)$就是 NB2 密度(20.12)，

其参数为 $\mu_j=\exp(\mathbf{x}'\boldsymbol{\beta}_j)$ 与 α_j，所以 $\boldsymbol{\theta}_j=(\boldsymbol{\beta}_j,\alpha_j)$。当成分数 m 是给定的时候，在某些正则条件下，对参数 $(\pi_j,\boldsymbol{\theta}_j)$ 进行极大似然估计是可行的，$j=1,2,\cdots,m$。

前面已经给出有限混合表示的优缺点，这里仅仅简要提及。在持续期限背景下的深入讨论则由 18.5 节给出。首先，有限混合是一种灵活又简约的数据建模方法。每一种混合成分都提供了对真实分布某一部分的一种局部近似。其次，有限混合方法具有半参数意义，因为它并不需要关于混合变量的任何分布假设。最后，在许多情况下，其结果都很容易进行解释。如果研究者特别地对源自公共政策观点的子总体行为感兴趣，那么有限混合表示就引人注目。倘若忽略潜类型，这样 $m=1$，则估计参数将是潜类型参数加权和。

另外，存在几个潜在困难。第一，我们很少具有关于设定成分个数的理论保证，而且如果一些成分不是充分不同的，那么确实不能区分它们。一种通常的做法是，以几个成分开始，然后增加一些成分，如果这样做，模型拟合显著地得到改进。在一些情况下，仅仅允许截距是各不相同的，而硬性规定，不同成分的斜率都是相等的。在这种过程中，必须小心谨慎，因为在 m 未知的情况下，并不完全知晓极大似然估计量的抽样性质。

有几个研究已表明，对于医疗保健的计数模型来说，有限混合模型拟合得相当好[德布和特里维迪(Deb and Trivedi，1997，2002)]。为此，一种可能的原因是，总体被个体潜健康状况分割。那些健康的人，或许大多数人，会产生低平均需求，然而那些有病的人会引致高平均需求。当被观测到的健康状况是不完全可观测时，有限分布模型可能会很好地分离子总体。

20.4.4 *截尾与删失*

在一些研究中，样本中要求包括从事关注活动的被抽样的个体。于是，计数数据是截尾的，因为数据仅仅在响应变量的某个范围内是可观测的。截尾计数的例子包括在调查期间每周搭乘公共汽车的汽车游行次数，在商业大街上的被抽样个体的购物次数，在所有这些情况下，我们不能观测到零计数，所以这种数据称为零截尾的(**zero-truncated**)，或更一般地，称为左截尾。右截尾是由大于某一个特定值而失去观测值引起的。

16.2 节已经给出，利用 ML 估计对截尾模型与删失模型的一般研究。这里专门研究计数数据。

截尾会导致非一致参数估计，除非对似然函数加以适当修改。考察零截尾情况。设 $f(y|\boldsymbol{\theta})$ 表示密度函数，而 $F(y|\boldsymbol{\theta})=\Pr[Y\leqslant y]$ 表示离散随机变量的累积分布函数，其中，$\boldsymbol{\theta}$ 表示参数向量。当小于正整数 1 的 y 的实现值都被省略时，得到零截尾密度为：

$$f(y|\boldsymbol{\theta},y\geqslant 1)=\frac{f(y|\boldsymbol{\theta})}{1-F(0|\boldsymbol{\theta})},\quad y=1,2,\cdots \tag{20.16}$$

这是专门研究零截尾泊松(**zero-truncated Poisson**)的情况，例如，$f(y|\mu,y\geqslant 1)=e^{-\mu}\mu^{y}/[y!(1-\exp(-\mu))]$。容易构造基于该密度的对数似然，从而获得极大似然估计值。

删失计数(**censored counts**)最普遍地是由计数汇总大于某一个值而产生的。当大于汇总值的总概率质量相对很小时,调查设计中经常这样做。截尾与删失之间的一个重要差异是,在删除情况下,对应于删失计数的协变量都是可观测的;在截尾情况下,计数结果既不是可观测的,协变量也不是可观测的。与截尾情况一样,如果错误使用了删失似然,删失会导致非一致的参数估计值。还可参见 16.2 节。

例如,大于某个已知值 c 的事件数被汇总成单一类别。从而,y 的某些值是不完全可观测的;其准确值是未知的,但知道它等于或大于 c。观测数据具有密度:

$$g(y|\boldsymbol{\theta})=\begin{cases}f(y|\boldsymbol{\theta}), & \text{当 } y<c\\ 1-F(c-1|\boldsymbol{\theta}), & \text{当 } y\geqslant c\end{cases} \tag{20.17}$$

其中,c 是已知的。

一种相对复杂的情况是样本选择(**sample selection**)内容[特泽(Terza, 1998)]。于是,仅当另一个与 y 潜在相关的随机变量大于某个门限值时,计数 y 才是可观测的。例如,为了见到医疗专家,人们首先必须看一般医师。

20.4.5 *修正计数模型*

引出本节修正计数模型的主要动因是,解决所谓的超额零(**excess zeros**)问题,数据存在的零比计数模型所预测的要多许多,诸如泊松计数模型,甚至是 NB2。

围栏模型或两部分模型

围栏模型(**hurdle model**)或两部分模型(**two part model**)(参见 16.4 节)放松了关于 0 与正整数均来自相同数据生成过程的假设。0 是由密度 $f_1(\cdot)$来决定的,所以 $\Pr[y=0]=f_1(0)$。正的计数来自截尾密度 $f_2(y|y>0)=f_2(y)/(1-f_2(0))$。为了确保概率和为 1,要用 $\Pr[y>0]=1-f_1(0)$去乘。因而:

$$g(y)=\begin{cases}f_1(0), & \text{当 } y=0\\ \dfrac{1-f_1(0)}{1-f_2(0)}f_2(y), & \text{当 } y\geqslant 1\end{cases} \tag{20.18}$$

只有 $f_1(\cdot)=f_2(\cdot)$时,才简化成标准模型。因而,在修正模型中,生成 0 与正计数的两种过程没有强制为相同的。尽管引出该模型的动因是研究超额零,但它还有能力对极少零问题进行建模。

围栏模型的极大似然估计涉及似然函数中的两项极大化:一个对应于 0 的,而另一个对应于正的。这样做简单易行。

围栏模型具有下述解释:它反映出两阶段决策过程。例如,病人开始找医生第一次出诊,但第二次或后来出诊则是由不同机制来决定的[波尔迈耶和乌尔里希(Pohlmeier and Ulrich, 1995)]。

回归应用使用了泊松模型或负二项式的围栏形式,这通过将 $f_1(\cdot)$与 $f_2(\cdot)$设定成前面给定的泊松或负二项式密度来获得。在一些应用中,对 0/1 结果进行建模的围栏部分中的协变量不需要与出现在截尾部分的那些协变量一样,尽管在实际应用中它们经常是相同的。围栏模型得到了广泛应用,而围栏负二项式模型则是相当灵活的。其缺点是该模型并不是非常简约,参数个数一般要加倍。而且,参数解释也不像没有同样的围栏模型那样容易。

在围栏设定中,对分布的选择至关重要。利用更灵活的分布给出负二项式的模型明显比泊松模型要有优势。围栏模型的条件均值是正概率与零截尾密度的条件均值的乘积。因此,当正确设定围栏模型时,利用泊松回归,蕴含着错误设定,从而导致非一致估计值。由条件均值设定的形式知,边际效应计算极为复杂,类似于16.4节使用的两部分模型。

含有零或零膨胀模型

第二种修正计数模型是含有零模型(**with-zeros model**)或零膨胀模型。这用具有密度 $f_1(\cdot)$ 的二值过程补充了计数密度 $f_2(\cdot)$。当二值过程以概率 $f_1(0)$ 取0值时,$y=0$。当二值过程以概率 $f_1(1)$ 取1值时,y 由计数密度 $f_2(\cdot)$ 取计数值0,1,2,…。这可通过两种方式设置零计数产生:当二值随机变量取1值时,一种是作为二值过程的实现值,而另一种则是作为计数过程的实现值。其密度是:

$$g(y)=\begin{cases} f_1(0)+(1-f_1(0))f_2(0), & \text{当 } y=0 \\ (1-f_1(0))f_2(y), & \text{当 } y\geqslant 1 \end{cases} \tag{20.19}$$

一些回归模型设 $f_1(\cdot)$ 是 logit 模型,而设 $f_2(\cdot)$ 是泊松或负二项式密度。这种模型与围栏模型相比使用很少。它具有对极少零进行建模的能力。

零膨胀计数模型在经济计量学中的应用,比其他统计学科的应用要少得多。

20.4.6 *离散选择模型*

计数数据可能在某些计数受限于类型数目分组之后,能够由离散选择模型方法来建模。例如,类型可以是0,1,2,3和4,类型数目也可能大于4更多。无序模型诸如15.4节曾讨论的多项式 logit 均不是简约的,而且更重要的是不适合。相反,应使用可辨别出数据顺序的时序模型。

一种此类模型是有序模型(**ordered model**)。这定义了一个不可观测潜变量 $y^*=\mathbf{x}'\boldsymbol{\beta}+u$,当 y^* 逐步超越较高门限时,$y=0,1,2,\cdots$ 的值就是可观测的,门限值也是待估参数。当 u 是 logistic 分布(或者标准正态分布)时,得到了有序 logit 模型或 probit 模型。当计数还可取负值时,有序模型(参见15.9节)特别有用,就如同当对净变化进行建模时所发生的,譬如对工业厂商数量的净变化。

另一种可能的时序模型,是通过设定关于 $\Pr[y=1\mid y\geqslant 0]$、$\Pr[y=2\mid y\geqslant 1]$ 等的二值模型序列来获得,尽管这显得繁琐。

最后,在一些情况下,除了计数,还可利用持续期限。例如,如果医生出诊日期是已知的,那么人们能对计数建模,比如说,月出诊次数或出诊的时间间隔期限。通常后一种方法更有效,因为它使用了更详细的数据,但计数回归还能提供有关协变量作用的有用信息[迪安和鲍尔肖(Dean and Balshaw, 1997)]。

20.5 部分参数模型

我们利用部分参数模型意指,关注通过条件均值和方差,甚至它们都不是完全设定的,来对数据加以建模。在20.5.1节,我们考察建立在条件均值与方差设定

基础上的一些模型。在 20.5.2 节，我们考察与评论最小二乘法的应用，而最小二乘法没有以显性方式对计数数据中的内生异方差性进行建模。在 20.5.3 节，将考察更多部分数的一些模型，诸如那些对条件均值给出不完全设定的模型。

该方法类似于 NLS，只是这里考虑到被建模为条件均值函数的异方差性。

20.5.1 *拟 ML 估计*

如同 20.2.1 节讨论的，当利用 PML 或 QML 时，估计量分布在比可导致特定似然函数的数据生成过程的假设更弱的假设条件下获得。

让我们重新考虑式(20.6)。给定关于 ω_i 的函数形式假设以及 ω_i 的一致估计值 $\hat{\omega}_i$，人们就能一致地估计出这种协方差矩阵。我们能使用泊松假设 $\omega_i=\mu_i$，但正如已注意到的，数据经常是过度分散的，$\omega_i>\mu_i$，一种普遍运用的方差函数是 $\omega_i=(1+\alpha\mu_i)\mu_i$，即在 20.4.2 节曾经讨论的 NB2 模型，还有 $\omega_i=(1+\alpha)\mu_i$，即 NB1 模型的方差函数。注意，在后者情况下，式(20.6)简化成 $V_{PML}\hat{\beta}_P]=(1+\alpha)(\sum_i\mu_i\mathbf{x}_i\mathbf{x}_i')^{-1}$，所以就过度分散($\alpha>0$)而言，由式(20.7)给出的通常 ML 方差矩阵低估了真实方差。

相反，若 $\omega_i=E[(y_i-\mathbf{x}_i'\boldsymbol{\beta})^2|\mathbf{x}_i]$是未设定的，则 $V_{PML}[\hat{\beta}_P]$的一致估计值可通过艾克一怀特(Eicker - White)稳健三明治方差估计公式适应这种情况获得。需要对式(20.6)的中间和式进行估计。当 $\hat{\mu}_i\xrightarrow{p}\mu_i$ 时，$N^{-1}\sum_i(y_i-\hat{\mu}_i)^2\mathbf{x}_i\mathbf{x}_i'\xrightarrow{p}\lim N^{-1}\sum_i\omega_i\mathbf{x}_i\mathbf{x}_i'$。因而，$V_{PML}[\hat{\beta}_P]$的一致估计值，通过用$(y_i-\hat{\mu}_i)^2$ 与 $\hat{\mu}_i$ 代替式(20.6)中的 ω_i 与 μ_i 得到。

当对方差函数形式存在疑惑时，建议利用 PML 估计量。从计算形式上讲，这在本质上与泊松 ML 的一样，其限制条件是该方差矩阵必须重新计算。对稳健方差的计算经常是标准软件包中的一个选项。

这些关于泊松 PML 估计的结果，在性质上类似于正态条件下线性模型的 PML 估计结果。这些结果可扩展到建立在线性指数家族密度基础之上的 PML 估计。在所有情况下，一致性仅仅要求对条件均值的正确设定[内尔德和韦德伯恩(Nelder and Wedderburn, 1972)，古里耶克斯等人(Gouriéroux et al., 1984a)]。这就产生了关于广义线性模型的大量统计文献[参见麦卡拉和内尔德(McCullagh and Nelder, 1989)]。这就允许有效推断提供条件值被正确设定以及将许多数据类型嵌套成特殊情况——连续(正态的)、计数(泊松)、离散(二项式)以及正的(伽玛)，如同 5.7.4 节所详细阐述的。许多复杂方法，诸如时间序列与面板数据模型，都在更一般的 GLM 框架下而不是特殊的关于计数数据框架下得到表述。

一些经济计量学家发现，一种更自然方式是，运用 GMM 框架而不是 GLM 框架。于是，起始点是条件矩 $E[(y_i-\exp(\mathbf{x}_i'\boldsymbol{\beta})|\mathbf{x}_i]=\mathbf{0}$。如果数据对于不同 i 而言是独立的，并且条件方差是均值的倍数，可以证明，最优工具选择是 $\mathbf{x}_i$，从而得到估计方程(20.5)；对于更详细内容，参见卡梅伦和特里维迪(Cameron and Trivedi, 1998, 第 37～44 页)。对于计数面板数据(参见 20.5.3 节)与内生回归元(**endogenous regressors**)，GMM 框架具有丰硕成果。关于计数的完全设定参数联立方程模型正处于其初期，所以工具变量方法引人注目。给定工具 $\mathbf{z}_i$，$\dim(\mathbf{z})\geq\dim(\mathbf{x})$，一旦满足 $E[(y_i-\exp(\mathbf{x}_i'\boldsymbol{\beta})|\mathbf{z}_i]=\mathbf{0}$，$\boldsymbol{\beta}$ 的一致估计量极小化：

$$Q(\boldsymbol{\beta})=\left[\sum_{i=1}^{N}(y_i-\exp(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{z}_i\right]'\mathbf{W}\left[\sum_{i=1}^{N}(y_i-\exp(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{z}_i\right] \quad (20.20)$$

其中,**W** 表示对称加权矩阵。

这种方法的优缺点如下。主要优点是,该方法做出很少的分布假设,因而避免了可能的模型错误设定。然而,对结果变量的离散性及其自然的异方差性的忽略,导致有效性的损失。对 **W** 矩阵的合适选择可缓解这一问题。进一步地,当较高阶矩潜在存在额外显著信息时,通过强调分布的一阶矩,IV 估计量对大数据的计数或许是敏感的。表 20.2 阐述了某些数据类型的特点,这些数据不便于利用 GMM 类型估计量进行模模。

20.5.2 *最小二乘法估计*

当关注于只对条件均值进行建模时,最小二乘法比上一节的方法更差。

当条件均值关于 $\mathbf{x}$ 是线性时,y 对 $\mathbf{x}$ 的线性最小二乘法回归(**Linear Least-squares regression**)就得出一致参数估计量。然而,对于计数数据,设定 $\mathrm{E}[y|\mathbf{x}]=\mathbf{x}'\boldsymbol{\beta}$ 是不合适的,因为它允许 $\mathrm{E}[y|\mathbf{x}]$ 出现负值。由于类似原因,线性概率模型对二值数据而言不合适。

人们可以考虑对 y 进行变换。特别地,对数变换 $\ln y$ 对 $\mathbf{x}$ 回归。当数据中包括 0 时,这个变换就会出现问题,如同通常情况那样。一种标准解决方式是添加一个常数项,比如 0.5,然后通过 OLS 对 $\ln(y+0.5)$进行建模。若我们对 $\mathrm{E}[y|\mathbf{x}]$而不是对 $\mathrm{E}[\ln y|\mathbf{x}]$ 感兴趣,则这个特定方法就引入了再变换的问题,参见毛拉(Mullahy, 1998)。然而,对线性模型的转换具有方便的优点,例如,当右边内生变量需要成为"工具"的时候,这样做就特别方便。

相反,一种更好的方式是使用含有指数均值设定的非线性最小二乘法;也就是说,估计非线性回归模型 $y=\exp(\mathbf{x}'\boldsymbol{\beta})+u$ 。重要的是,关于 NLS 估计量的统计推断是建立在艾克—怀特稳健标准误差的基础之上,因为该回归的误差项将是异方差的。

对计数而言,NLS 估计量通常比泊松伪 MLE 的有效性更差。NLS 的一阶条件是$\sum_i(y_i-\exp(\mathbf{x}_i'\boldsymbol{\beta}))\exp(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i=\mathbf{0}$。与泊松伪 MLE[参见式(20.5)]情况相比,它对残差进行加权。当 $\mathrm{V}[y_i|\mathbf{x}_i]$是 $\mathrm{E}[y_i|\mathbf{x}_i]$的倍数时,泊松伪 MLE 加权就是最优的。对于处理计数数据的内在异方差性来说,后者是一个更好的模型。

20.5.3 *半参数模型*

我们用半参数模型(**semiparametric models**)意指,部分参数模型具有无限维(成分)元素,如同 9.7 节所发展起来的。维数会激发我们对条件均值函数提出某种结构。

一类半参数模型就是不完全设定条件均值。重要的例子是,单指标模型与部分线性模型。单指标模型设定 $\mu_i=g(\mathbf{x}_i'\boldsymbol{\beta})$,其中,函数形式 $g(\cdot)$是未设定的,部分参数线性模型则设定 $\mu_i=\exp(\mathbf{x}_i'\boldsymbol{\beta}+g(\mathbf{z}_i))$,其中,函数形式 $g(\cdot)$未设定。在这两种情况下,在没有 $g(\cdot)$知识时,可获得 $\boldsymbol{\beta}$ 的$\sqrt{N}$一致渐近正态估计量。

第二个例子是，当假定 $\mu_i=\exp(\mathbf{x}_i'\boldsymbol{\beta})$ 而 $\mathrm{V}[y_i|\mathbf{x}_i]=\omega_i$ 是未设定时，对回归参数 $\boldsymbol{\beta}$ 进行最优估计。由于当 $N\to\infty$ 时会产生无限维元素，所以存在无限多个方差参数 ω_i。$\boldsymbol{\beta}$ 的最优估计量称为适应估计量，它就像知道 ω_i 一样有效。德尔加多和尼里斯纳(Delgado and Kniesner, 1997)利用核回归方法，对用于第二阶段非线性最小二乘法回归的权数加以估计，将线性回归模型的结果推广到含有指数条件均值函数的计数数据上。在他们的应用中，估计量几乎没有显示设定 $\omega_i=\mu_i(1+\alpha\mu_i)$ 的结果，并且 NB2 形式过度分散。

20.6 多变量计数与内生回归元

在本节，我们非常简略地介绍从横截面到计数数据的其他类型的推广[对于更详细的内容，参见卡梅伦和特里维迪(Cameron and Trivedi, 1998)]。对于多变量计数数据，可提出许多模型，但更受人们喜欢的方法尚未建立。对于面板数据，尽管统计学文献考察了较广泛的一些模型，但在经济计量文献中对使用哪一种方法却有更多的一致观点；参见 23.7 节。

20.6.1 *多变量数据*

在一些数据中，可以观测到比一个计数多的集合。例如，健康服务的几种不同类型的数据都可以得到利用，诸如医生出诊与住院天数。如果计数是相关的，联合建模将会改进有效性，并且提供数据更丰富的模型。本节将简要回顾和本章主要模型有关的二变量计数模型(**bivariate count models**)。熟悉含有相关误差项的多方程线性模型，比如 6.9.3 节的 SUR 模型(**SUR model**)，读者可考虑对含有相关误差项的多方程计数模型的推广。假定我们可观测到相同个体的几个计数变量(比如，看医生次数和拿处方药的次数。相关性来源会依赖于不可观测异质性。考虑相关误差联合估计将会产生更有效的估计值，但以额外计算复杂性为代价。

半参数方法

一旦将线性回归模型的一些方法适用到条件均值是非线性的且数据是异方差的计数数据上，部分参数方法则将这看成看似不相关回归问题；参见 6.10.3 节。

古里耶克斯、蒙福特和特罗格恩(Gouriéroux, Monfort and Trognon, 1984b)曾经提出基于矩方法推导二变量泊松类型模型。他们通过 y_1 与 y_2 的前二阶矩定义一个模型，然后通过准广义伪极大似然方法估计它。这种模型考虑到过度分散，而且它比二变量泊松模型更一般，但它却没有保持计数的整数值性质。

德尔加多(Delgado, 1992)将多变量计数模型看作多变量非线性模型，并提出半参数的广义最小二乘法估计量。利用 k - NN 方法对残差协方差矩阵加以估计。该方法不同于古里耶克斯、蒙福特和特罗格恩(Gouriéroux, Monfort and Trognon, 1984)在选择协方差矩阵估计量时的那种方法。

相当多的参数研究都使用两变量泊松模型。推导这种分布的一种方法是，假定当 $y_1=z_1+w$ 与 $y_2=z_2+w$(其中，所有 z_1、z_2 以及 w 都是独立的且服从泊松分布)时，就生成两个计数 y_1 与 y_2，正的参数 λ_1、λ_2 以及 λ_{12} 分别被参数化为外生协变

量的函数。这称为三变量归约(trivariate reduction)。

y_i 的边缘分布是泊松$[\lambda_j+\lambda_{12}]$,因此,这个模型将条件均值限定等于每一个计数变量的条件方差,所以:

$$\mathrm{E}[y_j|\mathbf{x}_j]=\mathrm{V}[y_j|\mathbf{x}_j] \tag{20.21}$$

对于 $j=1,2$,其中,$\mathbf{x}_j$ 表示解释变量的向量。相关系数由:

$$\mathrm{Cor}[y_1,y_2]=\frac{\lambda_{12}}{\sqrt{(\lambda_1+\lambda_{12})(\lambda_2+\lambda_{12})}} \tag{20.22}$$

给出,由于 $\lambda_{12}>0$,所以它是正的。

完全参数方法

对于每一个计数,通过引入不可观测异质性,最近几个研究发展了比较好的参数模型。有关问题已在 6.10.1 节与 19.3 节讨论过。

马歇尔和奥尔金(Marshall and Olkin, 1990)以下述方式考察两个计数边缘分布中含有乘法不可观测异质性(**multiplicative unobserved heterogeneity**)的模型。设 y_j 表示 $\mathcal{P}[\lambda_j\nu]$,$j=1,2$,其中,$\mathcal{P}$ 表示泊松分布,其均值为 $\lambda_j\nu$,而 ν 服从伽玛分布,其密度为:

$$g(\nu)=\frac{\nu^{\alpha-1}\exp(-\nu)}{\Gamma(\alpha)}$$

随机变量 ν 可以解释为共同(分享的)不可观测异质性。所得到的模型是一个因素模型(**one-factor model**)。两个计数的二变量负二项式(BVNB)分布被定义为:

$$\begin{aligned} f(y_1,y_2|\mathbf{x}_1,\mathbf{x}_2) &= \int_0^{\infty} f_1(y_1|\mathbf{x}_1,\nu)f_2(y_2|\mathbf{x}_2,\nu)g(\nu)d\nu \\ &= \int\Big[\prod_{j=1}^{2}\frac{\exp(-\lambda_j\nu)(\lambda_j\nu)^{y_j}}{y_j!}\Big]\frac{\nu^{\alpha-1}\exp(-\nu)}{\Gamma(\alpha)}d\nu \\ &= \frac{\Gamma(y_1+y_2+\alpha)}{y_1!y_2!\Gamma(\alpha)}\Big[\frac{\lambda_1}{\lambda_1+\lambda_2+1}\Big]^{y_1}\Big[\frac{\lambda_2}{\lambda_1+\lambda_2+1}\Big]^{y_2} \\ &\quad\times\Big[\frac{1}{\lambda_1+\lambda_2+1}\Big]^{\alpha} \end{aligned} \tag{20.23}$$

这种混合具有闭形式解,但该模型把不可观测异质性限定为两个计数变量的同一成分。联合似然组建成如同式(20.23)的项。其边缘分布都是单变量负二项式,而且两个计数变量之间的相关性

$$\mathrm{Cor}[y_1,y_2]=\frac{\lambda_1\lambda_2}{\sqrt{(\lambda_1^2+\alpha\lambda_1)(\lambda_2^2+\alpha\lambda_2)}} \tag{20.24}$$

必须是正的。

卡梅伦和约翰逊(Cameron and Johansson, 1998)、芒金和特里维迪(Munkin and Trivedi, 1999)以及奇布和温克尔曼(Chib and Winkelmann, 2001)提出了一些含有更灵活相关结构的模型,但在计算时要求高等方法。

芒金和特里维迪(Munkin and Trivedi, 1999)曾经考察如下 BVNB 模型的推广:

$$f(y_1,y_2\mid\mathbf{x}_1,\mathbf{x}_2)=\int_0^\infty\int_0^\infty f_1(y_1\mid\mathbf{x}_1,\nu_1)f_2(y_2\mid\mathbf{x}_2,\nu_2)g(\nu_1,\nu_2)d\nu_1 d\nu_2 \quad (20.25)$$

其中,联合分布是由两个边际模型组成的,每个模型以独立不可观测异质性变量为条件,ν_1 与 ν_2 分别被设定成二变量正态分布。以($\mathbf{x}_1,\mathbf{x}_2,\nu_1,\nu_2$)为条件的每一个边际模型都是含有乘法不可观测正态异质性的泊松模型。因此,该模型是二变量泊松对数正态混合(**bivariate Poisson-log-normal mixture**)。似然函数是在如同式(20.25)的样本上的乘积。作者将这解释成为"两因子模型"(**two-factor model**)。这种设定更具有灵活性,就如同它对两个不可观测成分之间相关性的符号与大小没有约束一样。然而,这种额外的灵活性引进了计算复杂性,因为式(20.25)中的二变量积分没有解析解,从而必须用基于模拟方法(第 12 章曾讨论)处理。当模型维数即 y 变量的个数增加时,所涉及的数值积分的阶数也会增加。结合了可能大样本量的这种特性使得计算任务非常繁重。奇布和温克尔曼(Chib and Winkelmann, 2001)提出一种可供选择的贝叶斯 MCMC 方法,该方法保留上述设定的灵活性,并处理了高维数结果向量。他们运用六维混合泊松对数正态模型阐述他们方法的灵活性。

最近发展起来的对相关计数进行建模的另一种方法是 19.3 节曾描述的联接方法。这里以对边缘分布进行设定开始;联合分布可利用联接,通过组合边缘分布获得。19.3 节曾给出相关持续期限的一些例子。还可参见卡梅伦、李、特里维迪和齐默(Cameron, Li, Trivedi, and Zimmor, 2004)。

20.6.2 *具有内生回归元的计数模型*

在许多背景下产生了计数变量的联立模型。例如,卡梅伦等人(Cameron et al., 1988)关注于计数变量(医疗使用),但协变量中的一个即健康保险状况主题是内生选择的。毛拉(Mullahy, 1997)在横截面背景下,而克雷蓬和迪盖特(Crépon and Duguet, 1997b)在面板数据背景下,将 GMM 方法应用到含有内生回归元的计数模型上。一个来自卫生经济学的非常著名的例子涉及医疗服务的计数模型,诸如就医次数,其中一个回归元是个体的医疗保险状况。对医疗保险的选择与结果方程误差项是不相关的假设是不现实的,因而保险回归元可能是内生的。第 22 章将提供更多例子,以及具有内生回归元的面板计数模型的详细内容。

当前的经济计量文献提供两种估计具有内生回归元的方法:一种是建立在 GMM/IV 基础上的方法,而另一种则是建立在极大似然较强假设基础上的方法。我们将依次考察它们。

第一种方法[毛拉(Mullahy, 1997)]以矩条件开始。考察含有可加零均值误差项的指数均值模型:

$$y_i=\mathrm{E}[y_i\mid\mathbf{x}_i]+\nu_i=\exp(\mathbf{x}_i'\boldsymbol{\beta})+\nu_i \quad (20.26)$$

$$\mathrm{E}[\nu_i\mid\mathbf{x}_i]\neq 0 \quad (20.27)$$

假定具有可利用的工具变量 $\mathbf{z}_i$,它满足矩条件:

$$\mathrm{E}[\nu_i\mid\mathbf{z}_i]=0 \quad (20.28)$$

$$\mathrm{E}[y_i-\exp(\mathbf{x}_i'\boldsymbol{\beta})\mid\mathbf{z}_i]=0$$

于是，假定存在足够多的可利用的矩条件，GMM 或非线性的工具变量估计就是可行的。这种方法已在 6.5.3 节讨论过。详细内容和有关讨论，读者可参考该节。不过注意到，在实施这个方法时，忽略变量计数性质，并且对模型进行处理就好像是对含有指数均值的任何其他非线性模型那样。另一方面，注意，异方差性非常可能与计数数据联系在一起，因而运用 GMM/工具变量方法时，应该考虑到这种复杂情况。

毛拉已经指出，乘法误差项设定具有某种优点。然而，这会产生不同矩条件。设：

$$\mathrm{E}[y_i \mid \mathbf{x}_i, \nu_i] = \exp(\mathbf{x}_i'\boldsymbol{\beta})\nu_i \tag{20.29}$$

从而得出矩条件：

$$\mathrm{E}\left[\frac{y_i}{\exp(\mathbf{x}_i'\boldsymbol{\beta})} - 1 \,\middle|\, \mathbf{z}_i\right] = 0 \tag{20.30}$$

它是 6.5 节曾讨论的非线性矩条件 $\mathrm{E}[r(y_i, \mathbf{x}_i, \boldsymbol{\beta}) \mid \mathbf{z}_i = 0$ 的一种特殊情况。倘若利用适当且充分的矩条件，则可运用 GMM 方法。可是，就计数变量而言，异方差性可能会再次出现，并且有效性会损失，因为忽略了变量的计数特性。

一种可供选择的方法是，联立处理因变量的计数特性，并且内生回归元问题是更为参数化的[特泽(Terza，1998)]。德布和特里维迪(Deb and Trivedi，2004)发展含有保险计划变量(D)作为回归元的计数(Y)与关于保险计划的二值选择模型的联合模型。在他们的模型中，内生性起因于结果(计数)方程与二值选择方程中存在的相关不可观测异质性。他们的模型具有下述结构：

$$\Pr[Y_i = y_i \mid \mathbf{x}_i, D_i, l_i] = f(\mathbf{x}_i'\boldsymbol{\beta} + \gamma_1 D_i + \lambda l_i) \tag{20.31}$$

$$\Pr[D_i = 1 \mid \mathbf{z}_i, l_i] = g(\mathbf{z}_i'\boldsymbol{\alpha} + \delta l_i) \tag{20.32}$$

其中，l_i 表示反映不可观测异质性的潜因素(**latent factors**)，而 δ 与 λ 表示有关的因子负荷[1](factor loadings)。以共同潜因素为条件的选择与结果变量的联合分布能写成：

$$\Pr[Y_i = y_i, D_i = 1 \mid \mathbf{x}_i, \mathbf{z}_i, l_i] = f(\mathbf{x}_i'\boldsymbol{\beta} + \gamma_1 d_i + \lambda l_i) g(\mathbf{z}_i'\boldsymbol{\alpha} + \delta l_i) \tag{20.33}$$

因为假定(Y，D)是条件独立的。

由于 l_i 是未知的，所以估计会出现此问题。尽管 l_i 是未知的，但假定 l_i 分布 h 是已知的，因此能对联合密度进行积分，即：

$$\Pr[Y_i = y_i, D_i = 1 \mid \mathbf{x}_i, \mathbf{z}_i] = \int [f(\mathbf{x}'_i\boldsymbol{\beta} + \gamma_1 D_i + \lambda l_i) g(\mathbf{z}'_i\boldsymbol{\alpha} + \delta l_i)] h(l_i) dl_i \tag{20.34}$$

一旦以这种形式计算，模型的未知参数可以通过最似然法得到估计。

为了简单起见，假定 $h(l_i)$ 没有未知参数。于是，极大似然估计量极大化联合

[1] 又称为因子输入，即原始变数与因子之间的相关系数。——译者注

似然函数 $L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \mid y_i, D_i, \mathbf{x}_i, \mathbf{z}_i)$，其中，$\boldsymbol{\theta}_1 = (\boldsymbol{\beta}, \gamma_1, \lambda)$ 与 $\boldsymbol{\theta}_2 = (\boldsymbol{\alpha}, \delta)$ 分别表示结果与计划选择方程中的参数，而 L 表示联合似然，其第 i 个元素是式(20.34)所定义的。为了识别，需要额外正规化约束。

给定关于 f、g 以及 h 的适当设定，估计的主要实际问题是，该积分通常没有闭形式解。MSL 估计量涉及用模拟样本类似形式(平均)代替期望，即：

$$\widetilde{\Pr}[Y_i = y_i, D_i = 1 \mid \mathbf{x}_i, \mathbf{z}_i] = \frac{1}{S}\sum_{s=1}^{S}[f(\mathbf{x}_i'\boldsymbol{\beta} + \gamma_1 D_i + \lambda\tilde{l}_{is})g(\mathbf{z}_i'\boldsymbol{\alpha} + \delta\tilde{l}_{is})] \tag{20.35}$$

其中，$\tilde{l}_{is}$ 表示来自密度 h 的伪随机数第 s 个采样(出自总数 S 个采样)，而 $\widetilde{\Pr}$ 表示模拟概率。于是，可定义数据的模拟似然函数。MSL 估计量对模拟对数似然求极大值。

在计数回归模型中，由内生虚拟回归元发展起来的这种方法能推广到多重虚拟以及多重结果上，不论是离散的还是连续的情况。其局限性是估计过程繁琐，它与工具变量类型估计量相比显得相当复杂。另外，如同任何联立方程模型一样，可识别性是一个问题。应用研究一般包括从 **x** 向量中排除 **z** 向量里的某些非平凡解释变量。

20.7 计数例子：进一步分析

现在，我们重新考察前面建立在泊松回归基础上、运用更为灵活的参数模型以 NB2 模型的分析。

20.3 节提供的表 20.5 中最后一列给出了 NB2 模型的一些结果。这里，还要报告稳健标准误差与 t 比率。首先，注意到，过度分散系数 α 是非常显著的。沃尔德检验统计量是 8.926，导致了对等分散性 ($\alpha=0$) 零假设拒绝决策。与此一致的是，对数似然从 −60.087 增大到 −42.777。很明显，模型拟合改进相当大。由于模型是嵌套的，所以没有必要报告 AIC 与 BIC。

表 20.6 的第 3 行给出来自 NB2 模型的预测频数。这些非常接近于观测到的频数，从而证实了模型拟合的改进，过度分散得到了解释。

然而，在可供选择的一些估计方法之间，系数本身看起来相当稳定，而且所有效应都得到准确测算，反映出大样本的特性。该结果的这些特性令人鼓舞，这显示 NB2 模型是合情合理的。正如由基本经济理论预测的那样，利用与共保率是负相关的。估计影响看起来对过度分散处理并不敏感。

另外，对建模进行精炼是可能的。例如，德布和特里维迪(Deb and Trivedi, 2002)将具有两种成分的有限混合模型与两部分(围栏)模型的效果进行比较，可以发现，后者拟合得更好。不过，围栏模型甚至比 NB2 模型拟合得要好。尽管这种精炼提供了额外信息，但这里给出的结果没有一个被认为会对利用价格敏感性基本问题引起误导。

对就医次数来说，NB2 模型发挥了更好的作用。然而，对于其他计数结果，甚至可能需要比 NB2 更灵活的模型。

20.8 应用研究

可以发现，那些非线性最小二乘法模型很容易利用关于泊松回归的软件包，这是一般的经济计量学和统计学软件包广泛利用的选项。要获得稳健的标准误差，需要小心谨慎。许多经济计量学软件包还包括负二项式回归与基本的面板数据模型。而一般的统计学软件则在广义线性模型模块里包含计数回归。标准软件包还会产生某种拟合优度的统计量，比如伪 R^2 测量，对于泊松模型来说，参见 8.7.1 节。

最近发展起来的一些模型，诸如有限混合模型、大多数时间序列模型以及动态面板数据模型，都需要发展各自特有的程序。一种有效的方法是，运用矩阵编程语言与以用户定义的目标函数为基础进行估计的软件结合起来。对于简单模型来说，许多计算机程序使得执行极大似然估计与(非常值得做的)对由用户定义的函数进行稳健方差估计成为可能。

除报告参数估计值之外拥有估计效果数量指标也很有用，如同 20.2.3 节所讨论的。像 20.2.4 节注意的，应小心谨慎，确保泊松回归模型所报告的标准误差以及 t 统计量，都建立在对过度分散而言稳健的方差估计值之上。

除估计之外，强烈建议设定检验用于评价估计模型的适宜性。就泊松横截面回归而言，很容易执行过度分散检验。对于任何参数模型，人们可比较计数的实际频率分布与拟合频率分布，尽管并不总是容易认识到，当观测计数的分布高度分散时，哪一个模型会失效。可以运用建立在实际频率与拟合频率基础上的正式统计设定和拟合优度。

在大多数实际应用中，人们可能面临模型选择问题。对于基于似然的未嵌入式模型来说，人们能使用选择准则比如赤池信息准则(AIC)，对于许多参数模型，AIC 建立在拟合对数似然基础上且有模型自由度的惩罚。

20.9 文献注释

20.2 卡梅伦和特里维迪(Cameron and Trivedi，1998)对本章涵盖的所有专题都曾给出更具体也更深入的研究，他们还提供了综合参考文献。温克尔曼(Winkelmann，1997)则提供了关于计数方面的经济计量学文献研究。统计学文献通常是在 GLM 背景下分析计数的。麦卡拉和内尔德(McCullagh and Nelder，1989)是标准的参考文献。经济计量学文献通常低估了 GLM 文献的贡献。法尔迈耶和塔茨(Fahrmeier and Tutz，1994)提供更多关于 GLM 的最近的经济计量解释。20.2 节的内容是标准的，并在许多地方出现。

20.3 德布和特里维迪(Deb and Trivedi，2002)给出 RHIE 数据的详细分析。

20.4 卡梅伦和特里维迪(Cameron and Trivedi，1986)提出负二项式的早期表示和应用。豪斯曼等人(Hausman et al.，1984)将该模型及其变形应用于面板

数据。对于 20.4.3 节的有限混合方法，参见德布和特里维迪(Deb and Trivedi, 1997)。关于 20.4.5 节的围栏模型应用，包括第一个提出该模型的毛拉(Mullahy, 1986)、波尔迈耶和乌尔里希(Pohlmeier and Ulrich, 1995)、古尔穆和特里维迪(Gurmu and Trivedi, 1996)。

20.5 古里耶克斯等人(Gouriéroux et al., 1984a, b)以及卡梅伦和特里维迪(Cameron and Trivedi, 1986)都详细阐述 20.5.1 节的伪 MLE。

20.6 20.6 节讨论的数据类型的回归模型处于发展初期。一个著名例外是(静态)面板数据计数模型已经很好地建立起来，其标准参考文献是豪斯曼等人(Hausman et al., 1984)。还可参见布伦奈斯和约翰逊(Brännäs and Johansson, 1996)。关于多变量计数数据的适当模型与含有内生回归元的模型研究是当前的一个活跃领域；参见特泽(Terza, 1998)以及德布和特里维迪(Deb and Trivedi 2004)。

习 题

20-1 假定 Y 表示泊松分布，均值为 μ。

(a) 验证前四阶矩分别是 μ、μ、μ 和 $3\mu^2+\mu$。

(b) 证明 $\Pr[Y=j]$ 与 $\Pr[Y=j-1]$ 之间存在线性关系，$j=1,2,\cdots$。

(c) 考虑含有 $\mu_i=\exp(\mathbf{x}_i'\boldsymbol{\beta})$ 回归的泊松 MLE。对泊松 MLE 方差的可能估计值包括 $\hat{\mathrm{V}}[\hat{\boldsymbol{\beta}}]=[\sum_i\hat{\mu}_i\mathbf{x}_i\mathbf{x}_i']^{-1}$ 与 $\tilde{\mathrm{V}}[\hat{\boldsymbol{\beta}}]=[\sum_i(y_i-\hat{\mu}_i)^2\mathbf{x}_i\mathbf{x}_i']^{-1}$。证明，倘若数据密度得到正确设定，则它们是渐近等价的(通过 N 标度)。

20-2 现在考虑泊松模型中的过度分散。

(a) 假定 $Y|\mu\sim\mathcal{P}[\mu]$，其中，$\mu=\exp(\beta_0+\beta_1x)$，$\beta_0=\gamma_0+\varepsilon$，而 ε 表示不可观测随机变量，满足 $\mathrm{E}[\varepsilon]=0$，$\mathrm{V}[\varepsilon]=\sigma^2>0$。证明，$\mathrm{V}[Y]>\mathrm{E}[Y]$。

(b) 考虑含有方差函数 $\mu+\alpha\mu^2$ 的 NB2 模型，概率质量函数已由式(20.12)给出。利用关于 $\alpha\in[0,3]$ 的四个不同值的图，描述 Y 的不同实现值的概率质量特性；在你的回答里，要强调靠近原点与位于右边尾部的函数特性。

(c) 对于 20.4.1 节中由式(20.12)给出的 NB2 密度，证明当 $\alpha\to0$ 时，该密度趋于泊松形式。[这可能是技巧性的。]

20-3 考虑含有条件均值 $\mu=\exp(\mathbf{x}'\boldsymbol{\beta})$ 的泊松回归模型。将估计问题看成未加权非线性平方问题，即 $y=\mathrm{E}[y|\mathbf{x}]+\varepsilon$，其中，$\mathrm{E}[y|\mathbf{x}]=\exp(\mathbf{x}'\boldsymbol{\beta})$，并且 $\varepsilon\sim\mathrm{iid}[0,\sigma^2]$。

(a) 推导关于$(\boldsymbol{\beta},\sigma^2)$的非线性最小二乘法估计。把关于 $\boldsymbol{\beta}$ 的最小二乘法与极大似然方程加以比较，并解释它们之间差异。

(b) 推导关于 $\boldsymbol{\beta}$的加权非线性最小二乘法估计。解释你对权数的选取。(权数常用于处理异方差性。)

(c) 将加权非线性最小二乘法与极大似然方程比较，如果可能，请解释其相似性。

20-4 考虑有限混合密度 $f(y|\boldsymbol{\theta})=\sum_{j=1}^{C}\pi_jf_j(y|\boldsymbol{\theta}_j)$、$C$ 个明显的潜类型或者

一些子总体的可加混合，含有未知的混合比 $\pi_1, \cdots, \pi_C$，其中，$\sum_{j=1}^{C}\pi_j=1$，$\pi_j>0$。这里，y 表示计数变量，而关于第 i 个观测值的第 j 个元素密度表示成：

$$f_j(y_i)=\frac{\Gamma(y_i+\psi_{ji})}{\Gamma(\psi_{ji})\Gamma(y_i+1)}\left(\frac{\psi_{ji}}{\lambda_{ji}+\psi_{ji}}\right)^{\psi_{ji}}\left(\frac{\lambda_{ji}}{\lambda_{ji}+\psi_{ji}}\right)^{y_i}$$

其中，$\lambda_{ji}=\exp(\mathbf{x}_i'\boldsymbol{\beta}_j)$，$\psi_{ji}=\lambda_{ji}^{k}/\alpha_j$，$\alpha_j>0$，而 $\boldsymbol{\theta}_j=(\boldsymbol{\beta}_j,\alpha_j)$。这里，$k$ 表示 0 或 1。该模型是含有 C 个元素的有限混合负二项式，并且当 $\alpha_j=0$ 时，专门研究有限混合泊松。

(a) 证明 $\mathrm{E}[y_i\,|\,\mathbf{x}_i]=\bar{\lambda}_i=\sum_{j=1}^{C}\pi_j\lambda_{ji}$，而 $\mathrm{V}(y_i\,|\,\mathbf{x}_i)=\sum_{j=1}^{C}\pi_j\lambda_{ji}^2[1+\alpha_j\lambda_{ji}^{-K}]+\bar{\lambda}_i-\bar{\lambda}_i^2$。

(b) 证明，仅仅建立在一阶矩基础上的任何混合模型均是不可识别的。

(c) 证明，建立在前二阶矩基础上的 C 个元素的混合泊松模型是可识别的。

20-5 [改编自巴尔塔基和李(Baltagi and Li, 1999)。]对由 20.2.4 节给出的泊松模型过度分散的一个简单检验是针对$[(y_i-\hat{\mu}_i)^2-y_i]/\hat{\mu}_i$ 对 $\hat{\mu}_i$ 回归中原假设零系数进行检验。[巴尔塔基和李(Baltagi and Li, 1999)]文献里，提出一种可供选择的检验涉及建立在$(y_i-\hat{\mu}_i)^2$ 对 $\hat{\mu}_i$ 回归的相同检验。后者由高斯—牛顿回归检验思想引发而形成(参见 10.3.9 节)。请分析这两种检验之间的差异，以及实施第二种检验方式差异的含义。

20-6 对于本题，请用本章数据的 50%子样本。

(a) 估计泊松回归与负二项式回归，其中，MDU 作为因变量，下述一些变量作为解释变量：LC、IDP、LINC、FEMALE、EDUDEC、XAGE、BLACK、HLTHG、HLTHE 和 HLTHP。完成下面原假设的似然比，即变量 LC 与 IDP 对 MDU 没有影响。

(b) 利用本章中满足 $g(\mu)=\mu$ 的方差公式(20.9)与满足 $g(\mu)=\mu^2$ 的公式(20.10)，检验泊松回归过度分散。数据更支持哪一个方差公式呢？你从这个习题得到什么结论？

(c) 估计负二项式模型(NB2)。将过度分散参数的估计值与(b)部分的估计值进行比较。请解释其相似点与不同点。

(d) 利用来自负二项式估计的结果，请比较处于良好健康(基准)的平均个体与处于不良健康(HLTHP=1)的平均个体，关于 CL 变动所估计的边际效应。

(e) 对于此泊松设定，估计由零部分(logit 或 probit)与正部分(在零点截尾的泊松)构成的“围栏形式”。将这些结果与那些来自正常泊松模型的结果进行比较。分析两种模型含义的相似点与不同点。依据你的分析，哪一个模型能更好地解释数据？

第五部分

面板数据模型

横截面模型具有某些内在局限性。它们主要是一些均衡模型，通常不能阐明事件跨时期相依性。横截面模型也没有令人满意地解决关于行为持久性来源的基本问题。这类持久性可能是行为方面的，即由真实状态相依性引起的；也可能是虚伪的，即总体不能控制的异质性行为的典型产物。由于面板数据也被称为纵向数据，包括相同对象周期性重复观测值，所以面板数据具有很大的潜力，用以解决横截面模型所不能满意处理的问题。第 21 章至第 23 章阐述面板数据的一些方法。就非线性面板数据模型而言，我们从第 21 章连续数据至第 23 章受限固变量进行系统研究。既考虑到固定效应模型，又顾及到随机效应模型。对于持久性专题，这三章在利用面板稳健推断方法中显得特别重要。

第 21 章回顾线性面板数据回归模型的重要且一般性结论，对于那些具有良好线性回归知识的读者来说，很容易阅读这一章，它不要求第二部分至第四部分所涵盖的内容。我们建议，对高等内容感兴趣的读者来说，首先应快速阅读本章内容，以便获得熟悉的重要概念和定义。

第 22 章是对第 21 章内容的重要推广，尤其考虑到当前变量马尔可夫相依性结构的动态面板。这种分析是置于 GMM 框架下展开的，这是目前本领域颇受多数应用者喜爱的方式。当这里的分析涉及众多详细内容时，就显得晦涩难懂。假如对 GMM 具有深刻认识，将有助于掌握本章主要结论。

第 21 章与第 22 章的结论，没有以一般而统一的方式推广到第 23 章非线性面板模型上。对于受限因变量面板模型来说，存在相对较少的一般性结论。尽管这样，我们仍在第 23 章以阐述某些一般性问题和方法开始。本章稍后几节将第四部分曾经研究过的横截面模型推广到面板数据内容上并加以讨论。这几节分别分析二值数据、计数数据、删失数据以及持续期限数据这四种类型模型，对于熟悉类似横截面模型的读者来说，这些应是为他们准备的通俗易懂的内容。

21
线性面板模型:基础

21.1 引 论

面板数据(**panel data**)意指对相同横截面在几个时期的重复观测值,尤其是对微观经济学应用中的个体或厂商而言。用于刻画此类数据的其他一些术语包括:纵向数据(**longitudinal data**)与重复测量(**repeated measures**)。关注内容是来自短面板(**short panel**)数据,意指对大量横截面个体观测几个时期,而不是长面板,例如,对很少的横截面单位观测众多时期。

面板数据的主要优点是,它提高了估计准确性。一旦对每个个体组合或混合几个时期的数据,这是增加观测值的结果。不过,为了进行有效统计推断,人们需要对给定个体时不同时期回归模型误差的可能相关性加以控制。特别地,在混合OLS回归中,一旦产生低估标准误差与很可能被扩大的 t 统计量,OLS标准误差一般公式典型地高估了准确性。

面板数据的第二个引人注目之处是,对固定效应模型进行一致估计的可能性,该模型考虑到可能与回归元相关的不可观测个体异质性。这种不可观测异质性导致省略变量偏倚,此偏倚在原则上通过利用仅仅单一横截面的工具变量方法加以修正,但在实际应用中,获得有效工具是困难的。如果假定不可观测的特定个体效应是可加的且时不变的,那么只含有两个时期的短面板数据给出了一种继续进行估计的方法。

除微观经济计量学之外,应用统计学的大多数学科都将不可观测个体异质性处理成与回归元独立的分布。于是,这种效应称为随机效应(**random effects**),尽管一个更好的术语是纯随机效应。与固定效应模型相比,这个较强假设具有允许一致估计所有参数包括时常值回归元系数的优点。可是,如果真实模型含有固定效应,那么随机效应与混合估计量都是非一致的。经济学家经常将随机效应模型的假设看成不是由数据所支持的。

面板数据的第三个引人注目之处是,运用它所得到的认识个体行为动态特性,比从单一横截面中所认识的要更多一些。因而,横截面可能会得出20%的贫困,但面板数据决定同样的20%是否每年处于贫穷之中。举一个有关例子,面板数据可确定个体工资或失业期长度的高序列相关性是否归因于个体拥有高薪水或长期

失业的特定意愿，或是否是拥有过去高薪水或失业的结果。该专题将推迟到第22章讨论。

若撇开固定效应是否是必需的基本问题不谈，线性面板数据模型及有关估计量从概念上看都是简单的。用于推导面板数据估计量的大量代数运算却并不顺应人们对基础的认识：面板数据估计量的统计性质会随假定模型以及它对不可观测效应的处理不同而变化。进一步地，绝大多数代数运算并不可以推广到非线性面板模型。

本章阐述各种线性面板数据模型的基本估计量。21.2节与21.3节分别深入详细地介绍广泛运用的模型及估计量，以及对每年工时与工资之间关系的应用。固定效应模型与随机效应模型之间的主要区别放在21.4节研究。21.5节至21.7节分别阐述混合模型、特定个体固定效应模型、特定个体随机效应估计方面的额外详细内容。21.8节考察线性面板数据模型的其他一些基本内容，比如推断与预测。

21.2 模型与估计量概览

面板数据提供了个体不同时间的行为和不同个体的信息。

甚至对于线性回归来说，与横截面数据情况相比，标准面板数据分析运用了更广泛的模型与估计量。几种标准模型由21.2.1节阐述，然后21.2.2节讨论几种估计量。表21.1给出一个总结，表明若数据生成过程是特定个体固定效应模型(**individual-specific fixed effects model**)，则几种估计量都是非一致的。

表21.1 线性面板模型，常见估计量与模型[a]

	假设模型		
	混合	随机效应	固定效应
β 估计量	(21.1)	(21.3)与(21.5)	只有(21.3)
混合 OLS (21.1)	一致的	一致的	非一致的
组间 (21.7)	一致的	一致的	非一致的
组内(或固定效应) (21.8)	一致的	一致的	一致的
一阶差分 (21.9)	一致的	一致的	一致的
随机效应 (21.10)	一致的	一致的	非一致的

[a] 此表只考虑 β。对于标准误差的正确计算，参见21.2.3节。

与横截面情况相比，求估计量的正确标准误差同样更为复杂。人们除需要控制可能的异方差性之外，还要控制给定个体时误差在不同时间上的相关性。21.2.3节将涵盖这个专题。

21.2.1 *面板数据模型*

一种相当一般的面板数据线性模型，允许其截距与斜率系数随个体和时间不同而变化，满足：

$$y_{it}=\alpha_{it}+\mathbf{x}_{it}'\boldsymbol{\beta}_{it}+u_{it}, \quad i=1,\cdots,N,\ t=1,\cdots,T$$

其中,y_{it}表示纯量因变量,$\mathbf{x}_{it}'$表示$K\times1$维自变量向量,u_{it}表示纯量扰动项,i表示横截面的个体(或厂商,或国家),而t表示时间。

由于待估参数比观测值更多一些,所以该模型太一般,从而不可估计。对α_{it}与$\boldsymbol{\beta}_{it}$随i及t而变化的程度以及误差u_{it}的特性,都需要施加进一步限制。

混合模式

最具约束性的模型是混合模型(**pooled model**),它设定常值系数,即横截面分析的通常假设,因此:

$$y_{it}=\alpha+\mathbf{x}_{it}'\boldsymbol{\beta}+u_{it} \tag{21.1}$$

若这个模型得到正确设定,并且回归元与误差项不相关,那么可利用混合 OLS 对其进行一致估计。然而,给定个体时误差项可能随时间变化是相关的,在此情况下,不应使用通常报告的标准误差,因为它们很可能是向下偏倚的。此外,采用固定效应模型(下面将定义)是适宜的,那么混合 OLS 估计量是非一致的。

个体及时间虚拟变量

模型(21.1)的一个简单变形,允许截距随不同个体与时间而变化,而其斜率参数却不变。于是,$y_{it}=\alpha_i+\gamma_t+\mathbf{x}_{it}'\boldsymbol{\beta}+u_{it}$,或者:

$$y_{it}=\sum_{j=1}^{N}\alpha_j d_{j,it}+\sum_{s=2}^{T}\gamma_s d_{s,it}+\mathbf{x}_{it}'\boldsymbol{\beta}+u_{it}\text{[1]} \tag{21.2}$$

其中,当$i=j$时,N个个体虚拟变量(**individual dummies**)$d_{j,it}$等于1,否则等于0;当$t=s$时,$(T-1)$个时间虚拟变量(**time dummies**)$d_{s,it}$等于1,否则等于0,并假定$\mathbf{x}_{it}$不包含截距。(若包含截距,则N个个体虚拟变量必须省掉一个。)

当$N\rightarrow\infty$且$T\rightarrow\infty$时,这种模型具有$N+(T-1)+\dim[\mathbf{x}]$个参数,它们能一致地得到估计。我们关注短面板(short panels),其中,$N\rightarrow\infty$,而T则不是。于是,γ_s能一致地得到估计,所以$(T-1)$个时间虚拟变量被直接并入回归元$\mathbf{x}_{it}$中。而其挑战在于一旦控制N个个体截距α_i,对参数$\boldsymbol{\beta}$进行估计。

不过,一种可能性是拥有观测值分组的虚拟变量,诸如因地区分组,在此情况下,第24章的集群方法(clustering methods)是有意义的。可是,这里我们对N个个体截距的全部集合N进行设定,从而导致$N\rightarrow\infty$时的问题。

固定效应与随机效应模型

特定个体效应模型允许每一个横截面单元拥有不同的截距,尽管所有斜率都是一样的,因此:

$$y_{it}=\alpha_i+\mathbf{x}_{it}'\boldsymbol{\beta}+\varepsilon_{it} \tag{21.3}$$

其中,ε_{it}对于不同i及t均是 iid 的。这是表述式(21.2)的一种更简单方式,这样做就将任何时间虚拟变量包括在回归元$\mathbf{x}_{it}$之中。α_i表示可捕获不可观测异质性的随机变量。18.2～18.5节以及20.4节对此已经研究过。

---

[1] 原著中该式等号右边缺少一项“u_{it}”,这里已经加上。—— 译者注

本章自始至终做出一个强外生性或严格外生性假设：

$$\mathrm{E}[\varepsilon_{it} | \alpha_i, \mathbf{x}_{i1}, \cdots, \mathbf{x}_{iT}] = 0, \quad t = 1, \cdots, T \tag{21.4}$$

因此，误差项被假定，以回归元的过去值、现在值以及未来值为条件的均值为 0。张伯伦（Chamberlain，1980）针对面板数据的外生性与外生性检验给出了详细讨论。强外生性剔除了含有滞后因变量的模型或含有作为回归元的内生变量的模型，对这些模型的讨论将推迟到第 22 章。

式（21.3）的一种变形是把 α_i 处理成不可观测随机变量，此随机变量潜在地与观测回归元 $\mathbf{x}_{it}$ 相关。这种变形称为固定效应（FE）模型，因为早期的研究均把这些效应建模成要估计的参数 $\alpha_1, \cdots, \alpha_N$。倘若出现固定效应，并与 $\mathbf{x}_{it}$ 相关，则许多估计量比如混合 OLS 都是非一致的。相反，在短面板数据中，为了确保对 $\boldsymbol{\beta}$ 的一致估计，就需要剔除 α_i 的可供选择的其他估计方法。式（21.3）的另一种变形假定不可观测个体效应 α_i 是随机变量，该随机变量服从与回归元独立的分布。这种模型称为随机效应（RE）模型[**random effects (RE) model**]，它通常做出另外的假设，即：

$$\begin{aligned} \alpha_i &\sim [\alpha, \sigma_\alpha^2] \\ \varepsilon_{it} &\sim [0, \sigma_\varepsilon^2] \end{aligned} \tag{21.5}$$

因而，既假定式（21.3）中随机效应是 iid 的，又假定式（21.3）中误差项是 iid 的。注意到，对式（21.5）没有设定什么特定的分布。对此模型而言，更准确的术语是单向特定个体随机效应模型或更简单的随机截距模型，以此区分含有更一般随机效应的模型，比如 22.8 节阐述的混合线性模型的模型。不过，另一个称谓则是随机成分模型。

固定效应术语会潜在地使人误导，而随机效应术语更为准确地体现出随机效应。为了避免这样混淆，李明宰（M-J. Lee，2002）把固定效应称为“有关效应”，而把随机效应称为“无关效应”。我们使用传统记号与术语，不过很明显，不论是在固定效应模型中，还是在随机效应模型中，α_i 是随机变量。

等相关性模型

可以把 RE 模型看成是混合模型的特殊化，因为 α_i 能被归入误差项中。于是，把式（21.3）看成是 y_{it} 对 $\mathbf{x}_{it}$ 的回归，其综合误差项 $u_{it} = \alpha_i + \varepsilon_{it}$ 以及式（21.5）蕴含：

$$\mathrm{Cov}[(\alpha_i + \varepsilon_{it}), (\alpha_i + \varepsilon_{is})] = \begin{cases} \sigma_\alpha^2, & t \neq s \\ \sigma_\alpha^2 + \sigma_\varepsilon^2, & t = s \end{cases} \tag{21.6}$$

因此，RE 模型利用了约束：综合误差 u_{it} 是等相关性（**equicorrelated**）的，因为对于 $t \neq s$，$\mathrm{Cor}[u_{it}, u_{is}] = \sigma_\alpha^2 / [\sigma_\alpha^2 + \sigma_\varepsilon^2]$ 并不随时间差分 $t - s$ 而变化。很显然，在 RE 模型中混合 OLS 将是一致的，却是无效的。随机效应模型，也被称为等相关性模型（**equicorrelated model**）或可交换误差模型（**exchangeable errors model**）。

固定效应模型与随机效应模型

其基本的差异在于模型有没有固定效应。现代经济计量学文献强调固定效应，但我们仍然要提供随机效应模型的详细情况。

一些作者，包括张伯伦（Chamberlain，1980，1984）以及伍德里奇（Wooldridge，2002），都在式(21.3)中使用等式：

$$y_{it}=c_i+\mathbf{x}_{it}'\boldsymbol{\beta}+\varepsilon_{it}$$

非常明显，这种个体效应不论在固定效应模型中还是在随机效应模型中均是随机变量。这两个模型都假定：

$$\mathrm{E}[y_{it}|c_i,\mathbf{x}_{it}]=c_i+\mathbf{x}_{it}'\boldsymbol{\beta}$$

特定个体效应 c_i 是未知的，而在短面板中不能得到一致估计，所以我们不能估计 $\mathrm{E}[y_{it}|c_i,\mathbf{x}_{it}]$。然后，我们能通过针对 c_i 取期望而剔除 c_i，得到：

$$\mathrm{E}[y_{it}|\mathbf{x}_{it}]=\mathrm{E}[c_i|\mathbf{x}_{it}]+\mathbf{x}_{it}'\boldsymbol{\beta}$$

对于 RE 模型，假定 $\mathrm{E}[c_i|\mathbf{x}_{it}]=\alpha$，所以 $\mathrm{E}[y_{it}|\mathbf{x}_{it}]=\alpha+\mathbf{x}_{it}'\boldsymbol{\beta}$，因此，不可能识别 $\mathrm{E}[y_{it}|\mathbf{x}_{it}]$。然而，在 FE 模型中，$\mathrm{E}[c_i|\mathbf{x}_{it}]$随 $\mathbf{x}_{it}$ 而变化，只是不知道是如何变化的，因此，我们不能识别 $\mathrm{E}[y_{it}|\mathbf{x}_{it}]$。不过，在具有短面板的 FE 模型中，一致估计 $\boldsymbol{\beta}$ 是可能的(如同下面将要讨论的)。因而，在 FE 模型中，尽管条件均值是不可识别的，但是识别边际效应

$$\boldsymbol{\beta}=\partial\mathrm{E}[y_{it}|c_i,\mathbf{x}_{it}]/\partial\mathbf{x}_{it}$$

却是可能的。例如，一旦控制个体效应，识别额外增加受教育年限的工资效应是可能的，即使个体效应与条件均值均是不可识别的。

在短面板中，FE 模型仅仅允许边际效应$\partial\mathrm{E}[y_{it}|c_i,\mathbf{x}_{it}]/\partial\mathbf{x}_{it}$的识别，以至于仅对时变回归元才可识别，所以例如种族或性别的边际效应是不可识别的。RE 模型允许对 $\boldsymbol{\beta}$ 的所有分量以及 $\mathrm{E}[y_{it}|\mathbf{x}_{it}]$的识别，但重要的 RE 假设为：$\mathrm{E}[c_i|\mathbf{x}_{it}]$是常值，在众多微观经济计量应用中被认为是站不住脚的。

21.2.2 *面板数据估计值*

现在，我们引进几个广泛使用的 $\boldsymbol{\beta}$ 面板数据估计量，进一步详细内容由21.5～21.7 节提供。这些估计量会在所用数据是横截面的还是时间序列变异程度方面不同，而它们的性质则会依照固定效应模型是否合适模型而变化。

回归元 x_{it} 可能是时常值的(**time-invariant**)，或者是时变的(**time-varying**)，满足 $x_{it}=x_i$ 对于 $t=1,\cdots,T$。对于一些估计量，尤其是下面所定义的组间估计量与一阶差分估计量，仅有时变回归元的系数是可识别的。

混合 OLS

混合 OLS 估计量可通过对不同 i 与 t 叠放成具有 NT 个观测值的长回归，并利用 OLS 进行估计

$$y_{it}=\alpha+\mathbf{x}_{it}'\boldsymbol{\beta}+u_{it},\quad t=1,\cdots,N,\ t=1,\cdots,T$$

而获得。当 $\mathrm{Cov}[u_{it},\mathbf{x}_{it}]=\mathbf{0}$ 时，为了一致性，要么 $N\rightarrow\infty$，要么 $T\rightarrow\infty$。

如果混合模型(21.1)合适，并且回归元与误差项不相关，那么很明显，混合 OLS 估计量是一致的。然而，建立在 iid 误差基础上的一般 OLS 方差矩阵此处是

不适宜的，因为给定个体时误差对不同的 t 几乎一定是正确相关的。NT 个相关观测值的信息就没有 NT 个独立观测值的多。

为了认识这种相关性，注意到，对于给定的个体，我们认为所有时间上的 y 具有很大相关性，所以 $\text{Cor}[y_{it}, y_{is}]$ 是很大的。甚至在包括一些回归元之后，$\text{Cor}[u_{it}, u_{is}]$可能仍是非零的，并且它经常是相当大的。例如，如果模型过高预测了一年的个体工资，那么它也可能过高预测了同一个体在其他年份的工资。RE 模型考虑到这种相关性，由式(21.6)知，对于 $t \neq s$，有 $\text{Cor}[u_{it}, u_{is}] = \sigma_\alpha^2 / [\sigma_\alpha^2 + \sigma_\varepsilon^2]$。

通常的 OLS 输出均把 T 个年份中的每一个处理成独立的信息，但是给定正误差相关性时，信息内容就比这要少得多。这会导致对估计量准确性的过高评估，认为它是非常高的，正如 21.3.3 节所阐明的，而正式证明由 21.5.4 节给出。因此，每当 OLS 应用于面板背景下，人们需要运用许多修正是可行的，这要依赖于相关性、对误差所假定的异方差性结构以及面板是短的还是长的(参见 21.5 节)。

如果真实模型是固定效应模型，那么混合 OLS 估计量是非一致的。为了理解这一点，把模型(21.3)重新写成：

$$y_{it} = \alpha + \mathbf{x}_{it}'\boldsymbol{\beta} + (\alpha_i - \alpha + \varepsilon_{it})$$

于是，如果个体效应 α_i 与回归元 $\mathbf{x}_{it}$是相关的，那么 y_{it} 对 x_{it} 的混合 OLS 回归及截距导致了 $\boldsymbol{\beta}$ 的非一致估计量，因为这类相关性蕴含着，综合误差项($\alpha_i - \alpha + \varepsilon_{it}$)与回归元相关。

总之，如果常值分数模型或随机效应模型是合适的，那么混合 OLS 就是适宜的，只是面板修正标准误差与 t 统计量必须用于统计推断中。如果固定效应模型是合适的，那么混合 OLS 是非一致的。

组间估计量

混合 OLS 估计量使用了既有时间变化又有横截面单位变化来估计 $\boldsymbol{\beta}$。

在短面板中，组间估计量只是使用横截面单位变化。以特定个体效应模型(21.3)开始。一旦对所有年份加以平均，得到 $\bar{y}_i = \alpha_i + \bar{\mathbf{x}}_i'\boldsymbol{\beta} + \bar{\varepsilon}_i$，这能重新写成组间模型：

$$\bar{y}_i = \alpha + \bar{\mathbf{x}}_i'\boldsymbol{\beta} + (\alpha_i - \alpha + \bar{\varepsilon}_i), \quad i = 1, \cdots, N \tag{21.7}$$

其中，$\bar{y}_i = T^{-1}\sum_t y_{it}$，$\bar{\varepsilon}_i = T^{-1}\sum_t \varepsilon_{it}$，而$\bar{\mathbf{x}}_i = T^{-1}\sum_t \varepsilon_{it}$。

组间估计量(**between estimator**)是出自 $\bar{y}_i$ 对截距及$\bar{\mathbf{x}}_i$ 回归的 OLS 估计量。它使用了各个不同个体之间的变化，并且是横截面回归的类似形式，即 $T=1$ 时的特殊情况。

如果式(21.7)中回归元$\bar{\mathbf{x}}_i$ 与综合误差($\alpha_i - \alpha + \bar{\varepsilon}_{it}$)是独立的，那么组间估计量是一致的。这将是常值系数模型与随机效应模型的情况。与之相比，对于固定效应模型来说，组间估计量是非一致的，因为 α_i 被假定成与 $\mathbf{x}_{it}$相关，从而与$\bar{\mathbf{x}}_i$ 相关。

组内估计量或固定效应估计量

组内估计量不同于混合 OLS 估计量或组间估计量，它探讨面板数据的特殊性。在短面板中，它测算了特定个体回归元与其时间均值离差和特定个体因变量与其时间均值离差之间的关系。这是利用不同时间上数据变化而完成的。

特别地,以特定个体效应模型(21.3)开始研究,该模型可嵌套在式(21.1)中,作为 $\alpha_i=\alpha$ 的特殊情况。于是,对时间加以平均,得到 $\bar{y}_i=\alpha_i+\bar{\mathbf{x}}_i'\boldsymbol{\beta}+\bar{\varepsilon}_i$。当消掉了 α_i 项,式(21.3)中的 y_{it} 去减这个均值,得出组内模型(**within model**):

$$y_{it}-\bar{y}_i=(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)'\boldsymbol{\beta}+(\varepsilon_{it}-\bar{\varepsilon}_i),\quad i=1,\cdots,N,\ t=1,\cdots,t \tag{21.8}$$

组内估计量(**within estimator**)是式(21.8)的 OLS 估计量。该估计量的特性是,在固定效应模型中它可以得到 $\boldsymbol{\beta}$ 的一致估计值,而混合 OLS 与组间估计量则不能。

由 21.6 节知,组内估计量具有几种解释。它称为固定效应估计量,因为如果 α_i 是固定效应且误差 ε_{it} 是 iid 的,那么它是 $\boldsymbol{\beta}$ 的有效估计量。本章关注于把固定效应处理成可以被忽略的冗余参数(**nuisance parameters**)的文献,因为关注内容只是对 $\boldsymbol{\beta}$ 的估计。相反,如果固定效应成为关注的内容,那么同样可以对它们加以估计。在短面板中,个体 α_i 的这些估计是非一致的,尽管就重要变量而言,它们的分布与其变化可能是有信息价值的。当 N 不是太大时,一种可供选择的且较简单的计算组内估计量的方法就是利用最小二乘法虚拟变量估计。不过,可能直接通过 y_{it} 对 x_{it} 与 N 个个体虚拟变量的 OLS 回归而直接估计,从而得出 $\boldsymbol{\beta}$ 的组内估计量和 N 个固定效应的估计值(参见 21.6.4 节)。组内估计量的另外一种解释是协方差估计量。最后,求特定个体的离差,等价于求 y_{it} 与 $\mathbf{x}_{it}$ 对个体虚拟变量辅助回归的残差,然后对残差进行研究。

组内估计的主要局限是,时常值回归元的系数在组内模型中是不可识别的。因为如果 $x_{it}=x_i$,那么 $\bar{x}_i=x_i$,所以 $(x_{it}-\bar{x}_i)=0$。许多研究都在探寻对时不变回归元效应进行估计。例如,在面板工资回归中,我们可能对性别或种族的效应感兴趣。因此,众多实践者倾向于不使用组内估计量。混合 OLS 估计量或者随机效应估计量允许对时常值回归元系数进行估计,只是当固定效应模型是正确模型时,随机效应估计量则是非一致的。

一阶差分估计量

一阶差分估计量同样利用了面板数据的特定。在短面板中,它测算了特定个体在回归元上单一时期的变化与特定个体在因变量上单一时期变化之间的联系。

具体地讲,以特定个体效应模型(21.3)开始研究。于是,一旦滞后单一时期,得出 $y_{i,t-1}=\alpha_i+\mathbf{x}_{i,t-1}'\boldsymbol{\beta}+\varepsilon_{i,t-1}$。当消掉了 α_i 项,由式(21.3)的 y_{it} 减去 $y_{i,t-1}$,得到一阶差分模型:

$$y_{it}-y_{i,t-1}=(\mathbf{x}_{it}-\mathbf{x}_{i,t-1})'\boldsymbol{\beta}+(\varepsilon_{it}-\varepsilon_{i,t-1}),\quad i=1,\cdots,N,\ t=2,\cdots,T \tag{21.9}$$

一阶差分估计量(**first-difference estimator**)是式(21.9)的 OLS 估计量。与组内估计量一样,此估计量在固定效应模型中会得到 $\boldsymbol{\beta}$ 的一致估计量,尽管时常值回归元的系数是不可识别的。若 ε_{it} 是 iid 的,则对于 $T>2$,一阶差分估计量就没有组内估计量更有效。

随机效应估计量

随机效应估计量也是探讨面板数据特性的估计量。

以特定个体效应模型(21.3)开始,但是假定随机效应模型如同式(21.5)一样,其中,α_i 与 ε_{it} 均是 iid 的。虽然混合 OLS 是一致的,但混合 GLS 将是更有效的。RE 模型的可行 GLS 估计量(参见 4.5.1 节)称为随机效应估计量,它可从变换模型

$$y_{it}-\hat{\lambda}\bar{y}_i=(1-\hat{\lambda})\mu+(\mathbf{x}_{it}-\hat{\lambda}\,\bar{\mathbf{x}}_i)'\boldsymbol{\beta}+v_{it} \tag{21.10}$$

的 OLS 估计中计算出,其中,$v_{it}=(1-\hat{\lambda})\alpha_i+(\varepsilon_{it}-\hat{\lambda}\bar{\varepsilon}_i)$是渐近 iid 的,而 $\hat{\lambda}$ 关于:

$$\lambda=1-\frac{\sigma_\varepsilon}{\sqrt{\sigma_\varepsilon^2+T\sigma_\alpha^2}} \tag{21.11}$$

是一致的。21.7 节提出了式(21.10)的推导,以及估计 σ_α^2 与 σ_ε^2 的方法,从而给出估计 λ 的方法。注意到,$\hat{\lambda}=0$ 对应于混合 OLS,$\hat{\lambda}=1$ 对应于组内估计,而且当 $T\to\infty$ 时,$\hat{\lambda}\to1$。这是 $\boldsymbol{\beta}$ 的两步估计量。

在 RE 模型条件下,RE 估计量是完全有效的,尽管其有效性提高与混合 OLS 相比不一定很大。然而,如果固定效应模型是正确模型,那么 RE 估计量是非一致的。

21.2.3 *面板稳健统计推断*

各种面板模型包括了一些误差项,这些误差项记为 u_{it}、ε_{it}和 α_i。在许多微观经济计量学应用中,有理由假定对于不同 i,误差具有独立性。然而,误差潜在是:(1) 序列相关的(给定 i 时对不同 t 而言是相关的);(2) 异方差性。有效统计推断要求对这两种因素加以控制。

4.4.5 节的怀特异方差一致估计量很容易被推广到短面板上,因为对于第 i 个观测值,当 $N\to\infty$,其误差方差矩阵具有有限维。因此,在没有假定个体内误差相关特定函数形式或异方差性条件下,可获得面板稳健标准误差。利用 GMM 的更有效估计量,则推迟到 22.2.7 节讨论。

一种至关重要的发现是,许多计算机软件包含面板命令,其计算的默认标准误差均假定 iid 模型误差,从而导致不正确推断。特别地,就 y_{it} 对 $\mathbf{x}_{it}$ 的混合 OLS 回归而言,在没有对个体效应进行控制时,很可能 $\mathrm{Cov}[u_{it},u_{is}]>0$,对于 $t\neq s$。一旦忽略这种相关性,非常可能导致低估标准误差,并且高估 t 统计量。正如 21.3 节阐明的数据例子,而 21.5.4 节在代数形式上证明了这一点。尽管误差序列相关包含固定特定个体效应或随机特定个体效应,并能得到简化,它却不能完全被剔除。此外,如同横截面数据通常所做的那样,人们需要控制潜在异方差性。

面板稳健三明治标准误差

在混合回归

$$\tilde{y}_{it}=\tilde{\mathbf{w}}_{it}\boldsymbol{\theta}+\tilde{u}_{it} \tag{21.12}$$

中,21.2.2 节的面板估计量可通过 $\boldsymbol{\theta}$ 的 OLS 估计来获得,其中,各种不同面板估计量对应于 y_{it}、$\mathbf{w}_{it}'=[1\quad \mathbf{x}_{it}']$、$u_{it}$ 的各种不同变换 $\tilde{y}_{it}$、$\tilde{\mathbf{w}}_{it}$、$\tilde{u}_{it}$。关键是,$\tilde{y}_{it}$ 仅仅是 $y_{i1},\cdots,y_{iT}$的已知函数,对于 $\tilde{\mathbf{w}}_{it}$ 与 $\tilde{u}_{it}$来说,有类似情况。

在混合 OLS 的最简单情况下，不必进行变换，而且 $\boldsymbol{\theta}=[\alpha\ \boldsymbol{\beta}']'$。就组内估计量而言，$\tilde{y}_{it}=y_{it}-\bar{y}_i$，$\tilde{\mathbf{w}}_{it}=(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)$，这里出现唯一时变回归元，并且 $\boldsymbol{\theta}$ 等于时变回归元的系数。就一阶差分估计而言，$\tilde{y}_{it}=y_{it}-y_{i,t-1}$，$\tilde{\mathbf{w}}_{it}=(\mathbf{x}_{it}-\mathbf{x}_{i,t-1})$，同时仅有时变回归元的系数是可识别的。对于随机效应，$\tilde{y}_{it}=y_{it}-\hat{\lambda}\bar{y}_i$，$\tilde{\mathbf{w}}'_{it}=(\mathbf{w}_{it}-\hat{\lambda}\bar{\mathbf{w}}_i)$，$\boldsymbol{\theta}=[\alpha\ \boldsymbol{\beta}']'$。这种变换能引起序列相关，即使基本误差是不相关的。

一种简便的方法是，对于给定个体时，对不同时期观测值进行叠放表示，得出：

$$\tilde{\mathbf{y}}_i=\widetilde{\mathbf{W}}_i\boldsymbol{\theta}+\tilde{\mathbf{u}}_i$$

其中，$\tilde{\mathbf{y}}_i$ 表示上面例子中的 $T\times1$ 维向量，只是就一阶差分模型而言，它表示 $(T-1)\times1$的，而 $\widetilde{\mathbf{W}}_i$ 表示 $T\times q$ 阶矩阵，或者就一阶差分模型而言，它表示 $(T-1)\times q$ 阶矩阵。进一步地，对 N 个不同个体进行叠放，得到：

$$\tilde{\mathbf{y}}=\widetilde{\mathbf{W}}\boldsymbol{\theta}+\tilde{\mathbf{u}}$$

因此，OLS 估计量的三种表达式是：

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\mathrm{OLS}}&=[\widetilde{\mathbf{W}}'\widetilde{\mathbf{W}}]^{-1}\widetilde{\mathbf{W}}'\tilde{\mathbf{y}}\\&=\Big[\sum_{i=1}^{N}\widetilde{\mathbf{W}}_i'\widetilde{\mathbf{W}}_i\Big]^{-1}\sum_i\widetilde{\mathbf{W}}_i'\tilde{\mathbf{y}}_i\\&=\Big[\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{\mathbf{w}}_{it}\tilde{\mathbf{w}}_{it}'\Big]^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{\mathbf{w}}_{it}\tilde{\mathbf{y}}_{it}\end{aligned}$$

其中，就一阶差分估计量而言，第三个等式是从 $t=2$ 到 T 求和，使用最方便的表达式将随着内容而变化。

为了考虑一致性，注意到如果模型被正确地设定，那么经过通常的代数运算，可得到 $\hat{\boldsymbol{\theta}}_{\mathrm{OLS}}=\boldsymbol{\theta}+[\widetilde{\mathbf{W}}'\widetilde{\mathbf{W}}]^{-1}\widetilde{\mathbf{W}}'\tilde{\mathbf{u}}$，或者：

$$\hat{\boldsymbol{\theta}}_{\mathrm{OLS}}=\boldsymbol{\theta}+\Big[\sum_{i=1}^{N}\widetilde{\mathbf{W}}_i'\widetilde{\mathbf{W}}_i\Big]^{-1}\sum_{i=1}^{N}\widetilde{\mathbf{W}}_i'\tilde{\mathbf{u}}_i$$

给定对不同 i 具有独立性，一致性的根本条件是 $\mathrm{E}[\widetilde{\mathbf{W}}_i'\tilde{\mathbf{u}}_i]=\mathbf{0}$。这经常需要比 $\mathrm{E}[u_{it}|\mathbf{w}_{it}]=0$ 更强的假设。充分假设是由式(21.4)给出的强外生性。例如，在比强外生性假设更强的假设下进行估计，会允许滞后因变量作为回归元，参见第 22 章。

于是，给定误差对于不同 i 具有独立性时，$\hat{\boldsymbol{\theta}}_{\mathrm{OLS}}$的渐近方差是：

$$\mathrm{V}[\hat{\boldsymbol{\theta}}_{\mathrm{OLS}}]=\Big[\sum_{i=1}^{N}\widetilde{\mathbf{W}}_i'\widetilde{\mathbf{W}}_i\Big]^{-1}\sum_{i=1}^{N}\widetilde{\mathbf{W}}_i'\mathrm{E}[\tilde{\mathbf{u}}_i\tilde{\mathbf{u}}_i'\,|\,\widetilde{\mathbf{W}}_i]\widetilde{\mathbf{W}}_i\Big[\sum_{i=1}^{N}\widetilde{\mathbf{W}}_i'\widetilde{\mathbf{W}}_i\Big]^{-1}$$

在这种面板设置背景下，$\mathrm{V}[\hat{\boldsymbol{\theta}}_{\mathrm{OLS}}]$的一致估计类似于获得 $\mathrm{V}[\hat{\boldsymbol{\theta}}_{\mathrm{OLS}}]$的一致估计的横截面问题，而该 $V[\hat{\boldsymbol{\theta}}_{\mathrm{OLS}}]$的一致估计对未知形式的异方差性而言是稳健的。唯一复杂情况是，出现向量 $\mathbf{u}_i$ 而不是出现纯量 u_i，倘若面板是短的，从而 $\mathbf{u}_i$ 维数有限，就不会产生问题。

这产生混合 OLS 估计量的渐近方差矩阵的面板稳健估计，这既控制序列相关性，又控制异方差性，它由

$$V[\hat{\boldsymbol{\theta}}_{\text{OLS}}]=\Big[\sum_{i=1}^{N}\widetilde{\mathbf{W}}_i'\widetilde{\mathbf{W}}_i\Big]^{-1}\sum_{i=1}^{N}\widetilde{\mathbf{W}}_i'\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'\widetilde{\mathbf{W}}_i\Big[\sum_{i=1}^{N}\widetilde{\mathbf{W}}_i'\widetilde{\mathbf{W}}_i\Big]^{-1} \tag{21.13}$$

给出，其中，$\hat{\mathbf{u}}_i=\hat{\mathbf{u}}_i=\tilde{\mathbf{y}}_i-\widetilde{\mathbf{W}}_i\hat{\boldsymbol{\theta}}$。对于短面板情况来说，式(21.13)中的估计量假定对于不同 i 具有独立性且 $N\to\infty$，否则允许 $V[u_{it}]$与 $\text{Cov}[u_{it},u_{is}]$随 t 和 s 而变化。其等价表达式是：

$$V[\hat{\boldsymbol{\theta}}_{\text{OLS}}]=\Big[\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{\mathbf{w}}_{it}\tilde{\mathbf{w}}_{it}'\Big]^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\tilde{\mathbf{w}}_{it}\tilde{\mathbf{w}}_{is}'\hat{u}_{it}\hat{u}_{is}\Big[\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{\mathbf{w}}_{it}\tilde{\mathbf{w}}_{it}'\Big]^{-1}$$

其中，$\hat{u}_{it}=\tilde{y}_{it}-\tilde{\mathbf{w}}_{it}'\hat{\boldsymbol{\theta}}$。这个估计量是由阿雷拉诺(Arellano，1987)针对固定效应估计量提出的。

如果命令具有聚集稳健标准误差选项(参见 24.5.2 节)，基于式(21.13)的面板稳健标准误差可通过常规的 OLS 命令计算出来。由于此聚集建立在个体上，所以对于个体 i 来说，选择合格者(**identifier**)作为聚集变量(**cluster variable**)。该方法用于获得由表 24.1 给出的面板稳健标准误差。"稳健"标准差术语能引起混淆。混合回归做出的一种普通误差是利用标准稳健标准误差选项(参见 4.4.5 节)OLS 回归(21.12)进行估计。不过，这仅对异方差性加以调整，而在实际应用中，在面板设置背景下，更为重要的是，对个体误差相关性进行修正。另一个普遍误差尽管具有较小影响，但它要使用假定同方差性的聚集稳健标准误差，从而 $E[\mathbf{u}_i\mathbf{u}_i']$ 对不同 i 而言是常值的。

面板自助法标准误差

自助法提供了一种可供选择的获得面板稳健标准误差的方法。其关键假设是，观测值对不同 i 而言是独立的，所以人们一定要执行自助序时程序，该程序对于 i 进行放回重新抽样，并且使用给定个体时的所有观测时期。对于数据 $\{(\mathbf{y}_i,\mathbf{X}_i),\ i=1,\cdots,N\}$，这会得到 B 个伪样本，而且对每个伪样本，人们实施 $\tilde{y}_{it}$ 对 $\tilde{\mathbf{w}}_{it}$ 的 OLS 回归，得出 B 个估计值 $\hat{\boldsymbol{\theta}}_b$，$b=1,\cdots,B$。于是，方差矩阵的面板自助法估计值为：

$$\hat{V}_{\text{Boot}}[\hat{\boldsymbol{\theta}}]=\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\boldsymbol{\theta}}_b-\bar{\hat{\boldsymbol{\theta}}})(\hat{\boldsymbol{\theta}}_b-\bar{\hat{\boldsymbol{\theta}}})' \tag{21.14}$$

其中，$\bar{\hat{\boldsymbol{\theta}}}=B^{-1}\sum_b\hat{\boldsymbol{\theta}}_b$。此自助法提供了没有渐近的精炼(参见 11.2.2 节)。给定对不同 i 而言的独立性，当 $N\to\infty$时，估计值是一致的。它在渐近形式上等价于估计值(21.13)，正如横截面情况下自助序对等价于怀特异方差一致估计一样。此自助法确实没有提供渐近精炼，尽管具有渐近精炼的自助法是可能的(参见 11.6.2 节)。

这一自助法能应用于依赖于对不同 i 而言具有独立性且 $N\to\infty$的任何面板估计量，包括 21.5.2 节短面板的混合可行 GLS 估计量。关键是要仅仅对不同 i 进行重新抽样，而不是既对不同 i 又对 t 进行重新抽样。

讨论

在个体层面上，对误差序列相关的标准误差进行修正的重要性不能过分强调。目前，计算机软件包确定没有自助执行这一点。伯特兰、杜弗洛以及马拉内森(Bertrand，Duflo，and Mullainathan，2004)在差异中的差分估计背景下(参见 22.6 节)，阐述了标准误差计算中向下偏倚的结果。他们发现，面板稳健方法与面

板自助法会很好地起作用，即使在应用中就州年份而言，数据 N（州的个数）相对很小，而渐近理论则使用 $N\rightarrow\infty$。

下述例子（参见 21.2 表）同样表明，对任何序列相关与自相关的标准误差进行修正的重要性。

21.3 线性面板例子：小时与工资

劳动经济学的一个重要课题是，劳动力供给对工资变化的响应。标准教科书的劳动供给模型提出，对于已经工作的人员来说，工资提高对劳动供给的效应是含混不清的，收入效应导致更少工作弥补了更多工作方向上的替代效应。

对成年男性进行横截面分析发现，对工时具有相对很小的反应。然而，一种可能情况是，这种关联是虚伪的，只是反映了不可观测工作意愿越强烈，正向联系的工资就越高。在不可观测工作意愿是时常值的假设下，面板数据分析就能控制这一点。例如，组内估计量通过测算超过平均水平（或低于平均水平）工资的时期中个体工作超过平均水平（或低于平均水平）时数的程度来完成此项任务。

数据是源自齐利亚克的 532 名男性 1979～1988 年 10 年期间的数据。关注变量是 lnhrs，即每年工时数的自然对数。单个解释变量是 lnwg，即小时工资的自然对数。我们考察回归模型：

$$\text{lnhrs}_{it}=\alpha_i+\beta\text{lnwg}_{it}+\varepsilon_{it}$$

其中，在某些模型中特定个体效应 α_i 简化成 α，而 β 测算了劳动供给的工资弹性。假定误差项 ε_{it} 对不同 i 而言是独立的，但给定 i 时它可能对不同 t 而言是相关的。正如提及的那样，我们期望劳动供给弹性 β 是小的且正的。

齐利亚克（Ziliak，1997）另外包括了年龄的二次项、孩子数以及有病的指示变量。对 β 的估计值及其标准误差而言，这些回归元与年份虚拟变量所得出的结果相对差异很小，为简单起见，这里对此省略。在第 22 章，我们将考察更一般模型，允许 lnwg 成为内生变量，同时允许 lnhrs 的滞后项出现在回归元中。

21.3.1 *数据概括*

对于 5 320 个观测值，lnhrs 与 lnwg 的样本均值分别为 7.66 与 2.61，蕴含着集合平均 2 120 小时以及每小时 13.60 美元。样本标准差分别为 0.29 与 0.43，这显示工资而非小时的百分比项拥有更大的变异性。

对于面板数据，知道变异性通常是否针对不同个体或不同时间而存在，非常有用。序列 x_{it} 围绕其总均值 $\bar{x}$ 的总变异能分解成：

$$\begin{aligned}\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\bar{x})^2 &= \sum_{i=1}^{N}\sum_{t=1}^{T}[(x_{it}-\bar{x}_i)+(\bar{x}_i-\bar{x})]^2\\ &= \sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\bar{x}_i)^2+\sum_{i=1}^{N}\sum_{t=1}^{T}(\bar{x}_i-\bar{x})^2\end{aligned}$$

因为向量积项之和为 0。总之，总平方和等于组内平方和加组间平方和。这产生了组内标准差 s_W 与组间标准差 s_B，其中：

$$s_W^2=\frac{1}{NT-N}\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\bar{x}_i)^2$$

以及：

$$s_B^2=\frac{1}{N-1}\sum_{i=1}^{N}(\bar{x}_i-\bar{x})^2$$

对于 lnhrs 与 lnwg 来说，它们的组内样本标准差与组间样本标准差分别是 0.22 和 0.18，以及 0.19 和 0.39。因此，工资总变异比工时总变异更大一些，这归因于个体变异比其工资的更大。对于组内个体变异，实际上工资变异稍微小于工时变异。

21.3.2 *面板数据估计量的比较*

表 21.2 概括了 21.2.2 节曾定义的标准面板估计量应用于这些数据的情况，还有三个不同的标准误差估计值。如同下面要详述的，统计推断应该要么使用面板稳健的标准误差，要么使用面板自助法标准误差。

表 21.2 小时与工资：标准线性面板模型估计量[a]

	POLS	组间	组内	一阶差分	RE-GLS	RE-MLE
α	7.442	7.483	7.220	0.001	7.346	7.346
β	0.083	0.067	0.168	0.109	0.119	0.120
稳健 se[b]	(0.030)	(0.024)	(0.085)	(0.084)	(0.051)	(0.052)
方根 se	[0.030]	[0.019]	[0.084]	[0.083]	[0.056]	[0.058]
默认 se	{0.009}	{0.020}	{0.019}	{0.021}	{0.014}	{0.014}
R^2	0.015	0.021	0.016	0.008	0.014	0.014
RMSE	0.283	0.177	0.233	0.296	0.233	0.233
RSS	427.225	0.363	259.398	417.944	288.860	288.612
TSS	433.831	17.015	263.677	420.223	293.023	292.773
σ_α	0.000		0.181		0.161	0.162
σ_ε	0.283		0.232		0.233	0.233
λ	0.000	—	1.000	—	0.585	0.586
N	5 320	532	5 320	4 788	5 320	5 320

[a] 列出 lnhrs 对 lnwg 的混合 OLS(POLS)、组间、组内、一阶差分、随机效应(RE)GLS 以及 MLE 线性回归。圆括号内数字为面板稳健的斜率系数的标准误差，方括号内数字为面板自助法的标准误差，大括号内数字为假定 iid 误差的默认估计值。R^2、均值平方误差方根(RMSE)、残差平方和(RSS)、总平方和(TSS)和样本量来自 21.2 节给出的适当回归。参数 λ 是式(21.11)后面定义的。

[b] se 表示标准误差。

斜率参数估计

斜率参数 β 的估计值对各种不同估计方法而言是不一样的。使用仅有横截面变分的中间估计小于混合 OLS 估计值。组内或者固定数应估计值 0.168 远远大于混合 OLS 估计值 0.083，同时利用 5%双尾检验统计，其标准误差估计为 0.084 与 0.085。一阶差分估计值也较大于混合 OLS 的估计值，只是一阶差分估计值相当小于组内估计值，这同样仅仅使用了时间原则变分，RE 估计值 0.119 或 0.120 位于组间估计值与组内估计值之间，这是人们所希望的，因为可以证明，RE 估计值

是组间估计值与组内估计值的加权平均。这两个 RE 估计值之间非常接近。而方差 σ_α^2 与 σ_ε^2 的估计量类似,导致了回归(21.10)中的相似值 $\hat{\lambda}=0.585$ 与 $\hat{\lambda}=0.586$。令人惊讶的是,RE 估计值有效性不如混合 OLS 估计值,这也预示 RE 模型对误差相关性建模失败。

哪一个估计值备受人们青睐呢?在所有模型(混合的、RE 以及 FE)中,组内估计量与一阶差分估计量都是一致的,而在固定效应模型下其他估计量是一致的。因此,最稳健的估计值是组内估计值 0.168 或一阶差分估计值 0.109。不过,利用这些更稳健的估计量会产生有效性损失,从组内标准误差 0.83 到一阶差分标准误差 0.85,都非常大于源自混合 OLS 与 RE 估计值的那些误差。正式豪斯曼检验(参见 21.4.3 节的详细内容及讨论)能用于检验个体效应是否是固定的。考虑到此例估计相对含糊不清,尽管 FE 估计与 RE 估计之间差异很大,但豪斯曼检验没有拒绝随机效应的零假设。因而,这里可使用更有效的随机效应估计。随机效应估计的另一个优点是,它允许对时常值估计量的系数进行估计。

标准误差估计

现在,我们转到标准误差估计的比较上,由表 21.2.3 知,建立在面板稳健标准误差的基础之上,该标准误差允许对给定个体而言不同时期的误差是相关的,同时拥有随不同个体而变化的方差与协方差。同样地,正如后面几节所阐述的,为了解释损失 $N+K$ 个而不是 K 个自由度,需要估计量建立在平均偏差诸如式(21.8)与式(21.10)基础上的标准误差。

第一个标准误差估计是通过由式(21.13)给出的面板稳健方法计算,而第二个标准误差则是通过由式(21.14)给出的具有 500 次复制的面板自助法进行计算。为了简洁起见,这些估计值称为面板稳健的,尽管它们对异方差性同样是稳健的。这两个估计值非常接近,除了随机效应模型中的面板稳健标准误差被低估之外,这是因为它们是用回归(21.10)来进行计算的,计算中忽略了 $\hat{\lambda}$ 上的估计误差。

第三个标准误差估计是标准默认的计算机输出,这样的输出是建立在 iid 误差假设之上。在此例中,正确估计的标准误差显著地是默认标准误差的 3~4 倍。一个例外是组间估计量,由于它仅仅使用横截面变异,所以该估计量具有只需对异方差性加以修正的标准误差。

例如,对于 β 的混合 OLS 估计量,其默认标准误差是 0.09,得到不正确的 t 统计量 9.07。面板稳健标准误差非常大,为 0.30,得到的正确 t 统计量则相当小,为 2.83。默认标准误差假定对于给定 i 时模型误差对不同 t 具有独立性,可是时间上它们可能正相关。这种错误假设高估了其他时期的好处,从而得到标准误差向下偏倚(参见 21.5.4 节)。另外,忽略误差上的异方差性同样会导致偏倚,尽管此偏倚位于两者之中的任一方向上。对于这些数据来说,控制异方差性失败也会给予大的向下偏倚:控制异方差性而不是对给定 i 时不同 t 的相关性的 $\hat{\beta}_{\text{POLS}}$ 标准误差是 0.020。对于其他数据来说,对异常差性的修正通常没有对面板相关性修正那样重要。

对于组内估计量与组间估计量,包括 α_i 项,应该对给定个体控制不同时间误差上的某种相关性。不过,就这些数据而言,面板稳健标准误差与非稳健标准误差

之间的差异仍然很大，部分归因于额外控制异方差性的失败。

很明显，应该使用面板稳健标准误差。

21.3.3 图形分析

进行回归、组间回归以及固定效应（组内或一阶差分回归）的图形比较是一种有深刻见解的方式。尽管这样的图形在面板数据回归中很少画出，但是它们极容易应用在这里，因为仅存在一个回归元。

全部图形包括利用 Lowess 光滑元（参见 9.6 节）的非参数回归与由表 21.2 给出估计值的线性回归。图 21.1 画出，全部年份中所有厂商（5.320 观测值）的 lnhrs 对 lnwg 的图形。该图显示正的关系，除端点之外大致为线性的，而且由表 21.2 知，此线斜率为 0.083，具有很小的 $R^2=0.015$。

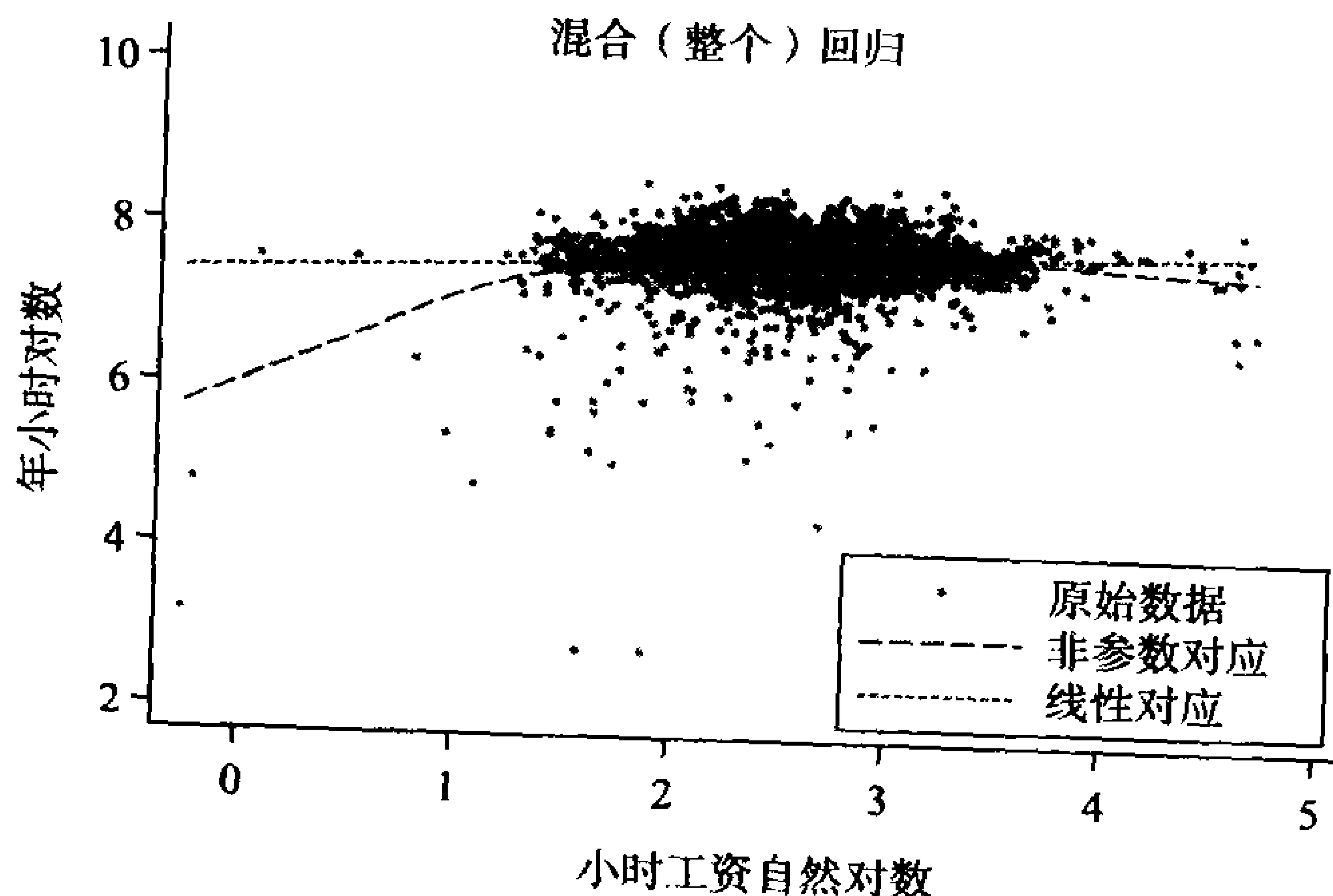


图 21.1 小时与工资：混合（整个）回归。画出年度工时自然对数对应小时工资自然对数的图形。数据是 1979～1988 年的 10 年期间每年 532 个美国男性。

组间估计量(21.7)是 $\bar{y}_i$ 与 $\bar{x}_i$ 进行回归。其相应的关于 lnhrs - lnwg 的数据图形，已由图 21.2 给出，再次表明正的关系。

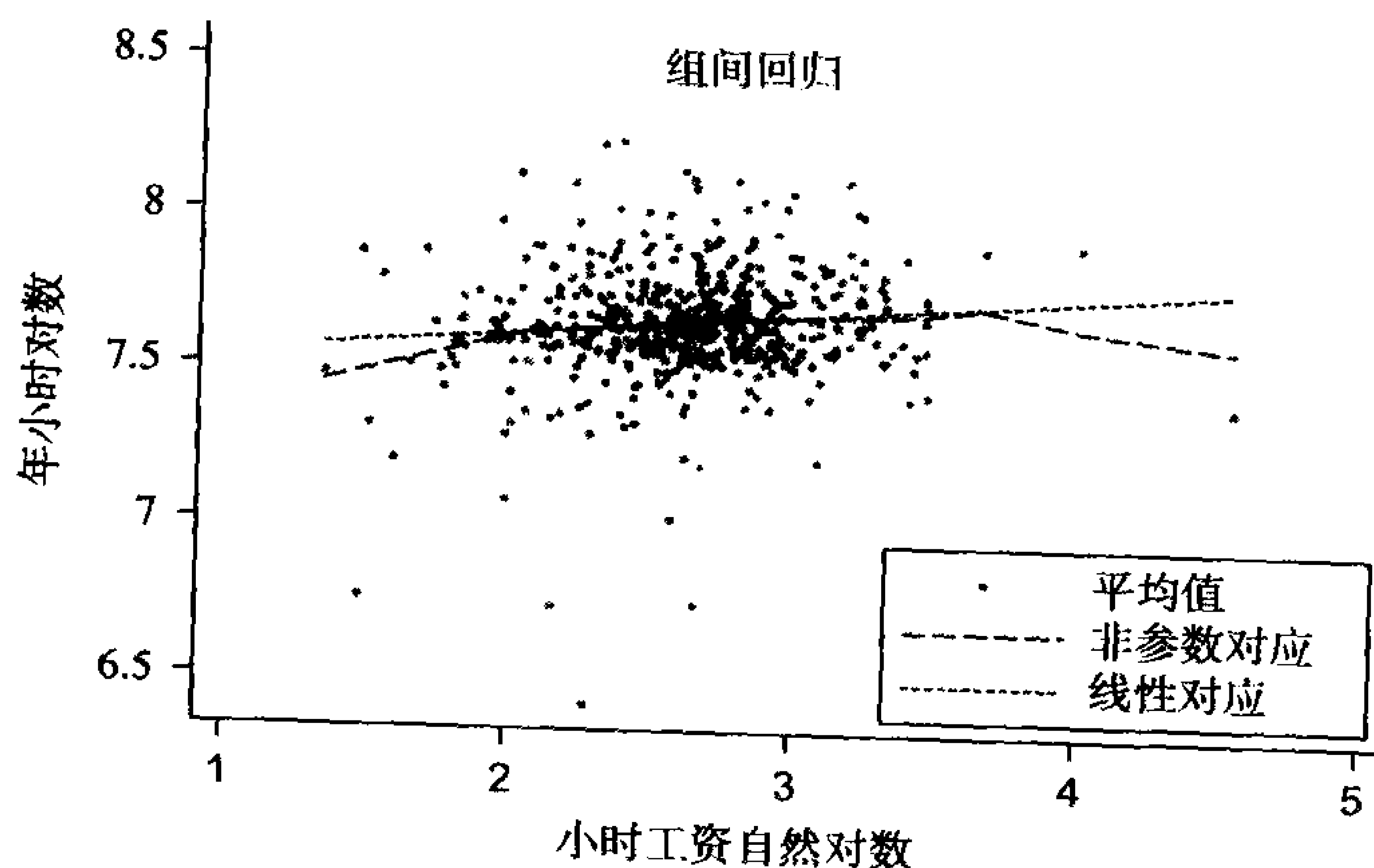


图 21.2 小时与工资：组间回归。画出工时自然对数的 10 年平均值对应 532 个小时工资自然对数的 10 年平均值的图形。样本与图 21.1 的一样。

组内或固定效应估计量(21.8)是$(y_{it}-\bar{y}_i)$对$(x_{it}-\bar{x}_i)$进行回归。图 21.3 给出了有关$(y_{it}-\bar{y}_i+\bar{y})$对$(x_{it}-\bar{x}_i+\bar{x})$回归的图形,其中,$\bar{y}=N^{-1}\sum_i\bar{y}_i$ 与 $\bar{x}=N^{-1}\sum_i\bar{x}_i$ 表示 y 与 x 的总均值。与图 21.1 相比,它表明对个体值进行差分,极大地促使 lnwg 变异性范围减少,而 lnhrs 的变异性减少却并不大。与混合 OLS 情况相比,其斜率表现得更加陡峭,并由表 21.2 知,其斜率从 0.083 增大到 0.168。

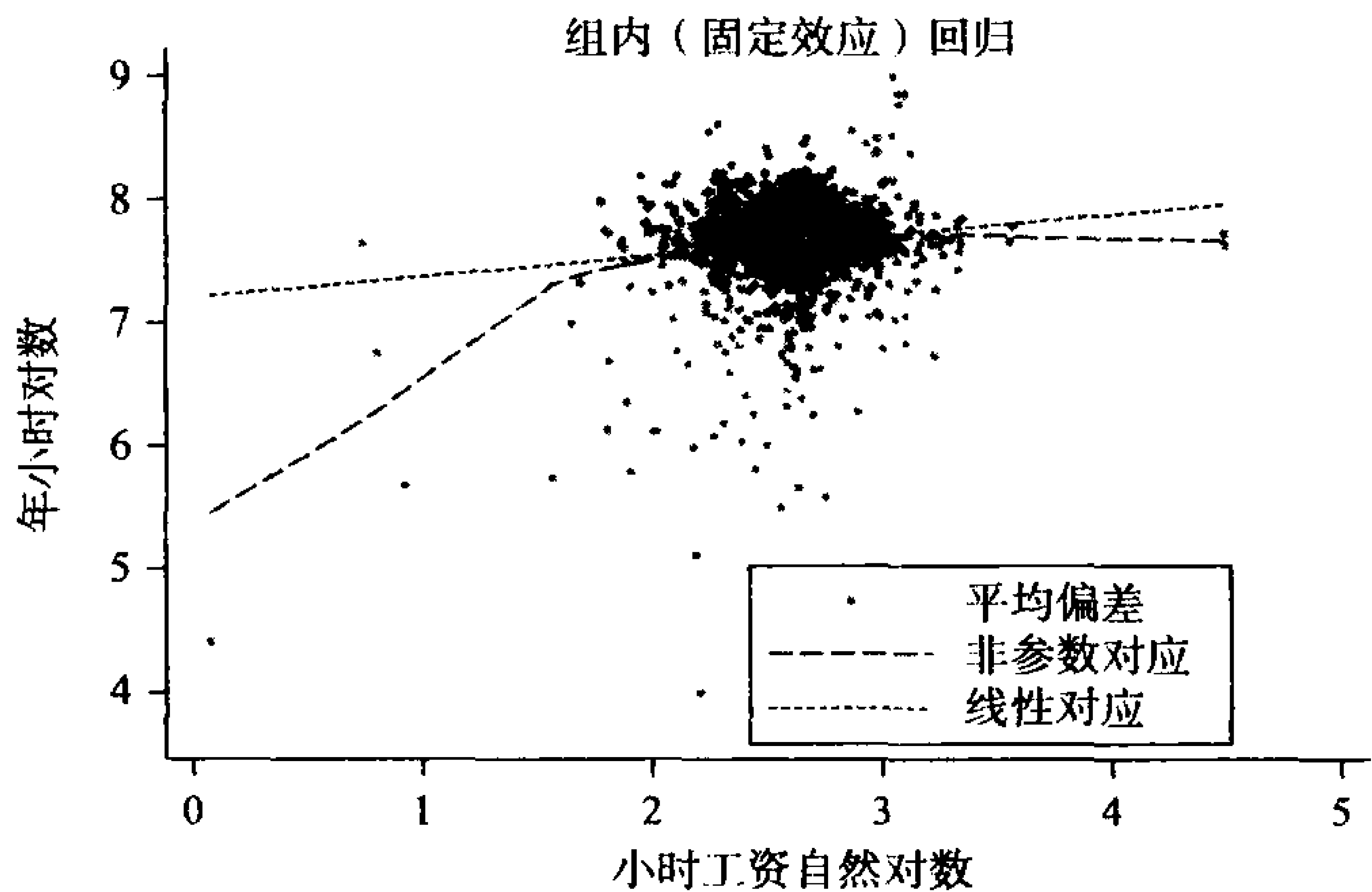


图 21.3 小时与工资:组内(固定效应)回归。利用 532 名男性 10 年数据画出工时自然对数 10 年平均偏差对应小时工资自然对数 10 年平均偏差图形。样本与图 21.1 的一样。

一阶差分估计量(21.9)是$(y_{it}-y_{i,t-1})$对$(x_{it}-x_{i,t-1})$进行回归。其关于 lnhrs - lnwg 数据的相应图形由图 21.4 给出。其性质上类似于图 21.3。

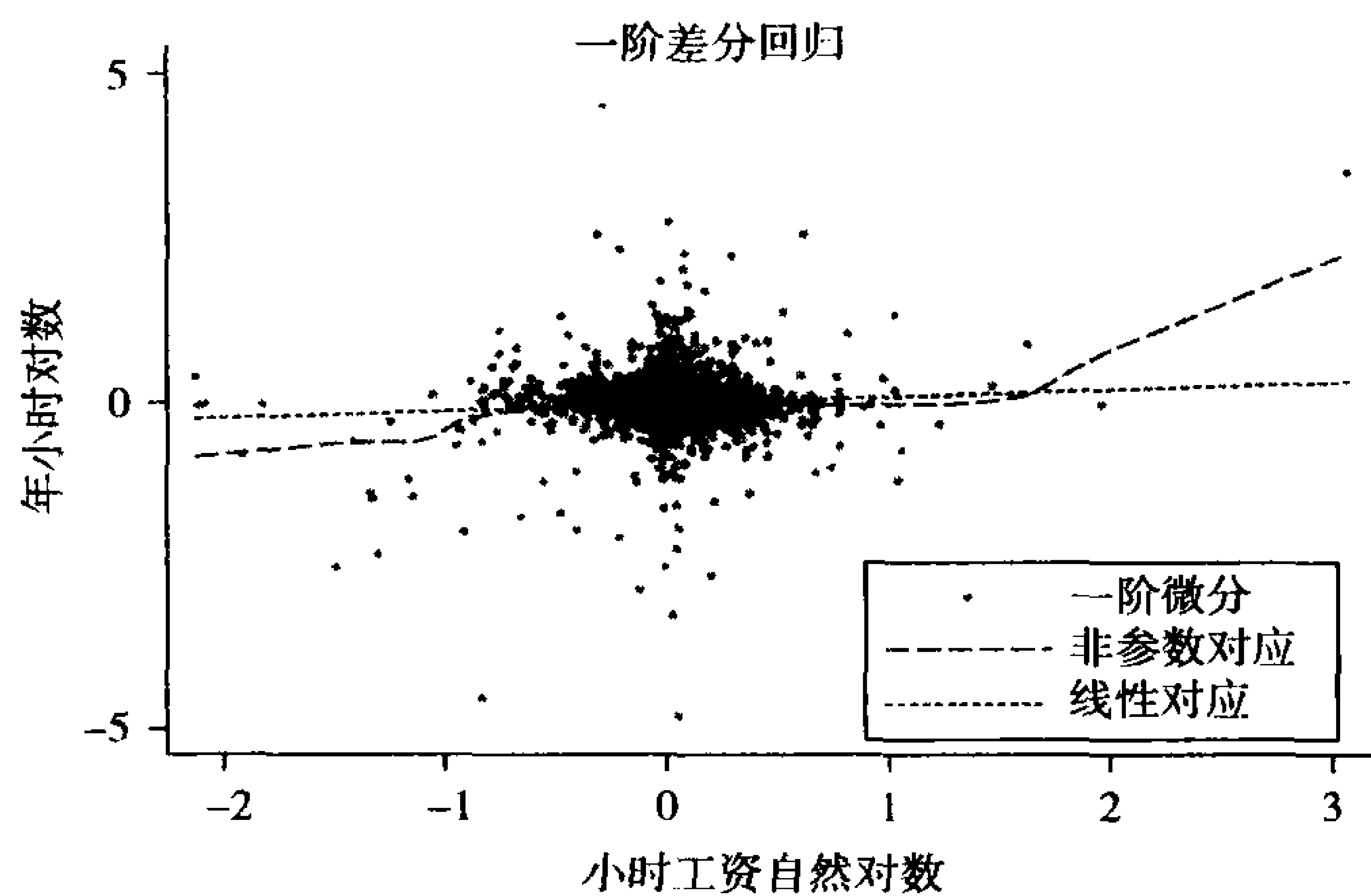


图 21.4 小时与工资:一阶差分回归。利用 53 名男性 10 年数据画出工时自然对数的一阶差分对应小时工资自然对数一阶差分图。样本与图 21.1 的一样。

前面分析的结论是,利用时间序列变异得出的对工资变化的响应,比利用横截面变异性所得到的对工资变化的响应要大一些。

21.3.4 *残差分析*

考察数据与残差的自相关模式具有意义。例如,对于残差 $\hat{u}_{it}=y_{it}-\hat{y}_{it}$ 来说,其在时期 s 与时期 t 之间的自相关计算为 $\hat{\rho}_{st}=c_{st}/\sqrt{c_{ss}c_{tt}}$,$s,t=1,\cdots,T$,其中,协方

差估计 $c_{st}=(N-1)^{-1}\sum_i(\hat{u}_{it}-\bar{\hat{u}}_t)(\hat{u}_{is}-\bar{\hat{u}}_s)$，而 $\bar{\hat{u}}_t=N^{-1}\sum_i\hat{u}_{it}$。

表 21.3 给出 lnhrs 对 lnwg 的混合 OLS 回归之后的残差自相关。关于 2～9 个时期的各个自相关通常位于 0.2 与 0.4 之间。衰变速率非常慢，而且自相关表现出更接近于随机效应模型，该模型假定与具有指数衰变的 AR(1) 相比，$\mathrm{Cor}[u_{it},u_{is}]$为常值，对于 $t\neq s$。

表 21.3　小时与工资：混合 OLS 残差的自相关[a]

	u79	u80	u81	u82	u83	u84	u85	u86	u87	u88
ufe79	1.00									
ufe80	0.33	1.00								
ufe81	0.44	0.40	1.00							
ufe82	0.30	0.31	0.57	1.00						
ufe83	0.21	0.23	0.37	0.47	1.00					
ufe84	0.20	0.23	0.32	0.34	0.64	1.00				
ufe85	0.24	0.32	0.41	0.35	0.38	0.58	1.00			
ufe86	0.20	0.19	0.28	0.25	0.31	0.35	0.40	1.00		
ufe87	0.20	0.32	0.33	0.29	0.31	0.34	0.39	0.35	1.00	
ufe88	0.16	0.25	0.30	0.26	0.21	0.25	0.34	0.55	0.53	1.00

[a] 注意：残差自回归是来自 532 名男性 10 年期间的 lnhrs 与 lnwg 的混合 OLS 回归。此自回归缓慢变弱。

关于回归前的 lnhrs 相关非常接近于那些由表 21.3 给出的情况，因为 $\hat{u}_{it}\simeq y_{it}$ 作为源自具有 $R^2=0.015$ 的混合 OLS 的不好的解释证据。虽然关于回归元 lnwg 的自相关在这里没有画出，但它更大一些，其范围大致从滞后一期的 0.9 到滞后 9 期的 0.7。

源自组内回归残差自相关已由表 21.4 给出。如果最初式(21.3)中误差 ε_{it} 是 iid 的，那么可以证明，变化误差 $\varepsilon_{it}-\bar{\varepsilon}_i$ 在所有滞后上具有等于 $-1/(T-1)=-0.11$ 的自相关。有一些违背这里的情况，尤其是对于第一滞后期来说，它总是正的。

表 21.4　小时与工资：组内回归残差自相关[a]

	u79	u80	u81	u82	u83	u84	u85	u86	u87	u88
ufe79	1.00									
ufe80	0.10	1.00								
ufe81	0.21	0.08	1.00							
ufe82	0.00	−0.04	0.26	1.00						
ufe83	−0.26	−0.27	−0.21	0.01	1.00					
ufe84	−0.26	−0.27	−0.30	−0.20	0.32	1.00				
ufe85	−0.18	−0.10	−0.11	−0.17	−0.16	0.17	1.00			
ufe86	−0.19	−0.25	−0.26	−0.27	−0.17	−0.14	−0.08	1.00		
ufe87	−0.15	−0.05	−0.16	−0.20	−0.24	−0.21	−0.09	−0.09	1.00	
ufe88	−0.17	−0.11	−0.14	−0.18	−0.38	−0.31	0.13	0.24	0.24	1.00

[a] 残差自相关来自 532 名男性 10 年期间的 lnhrs 与 lnwg 组内(固定效应)回归。

源自随机效应回归残差自相关相当类似于由表 21.4 给出的那些固定效应情形。源自一阶差分回归的残差自相关在性质上类似于以下理论结果：如果最初式

(21.3)中误差是 iid 的，那么变换误差 $\varepsilon_{it}-\varepsilon_{it-1}$ 在滞后一个时期具有自相关 0.5，而在其他滞后时期自相关为 0。

21.4 固定效应与随机效应模型

固定效应模型的优点是，允许研究者使用面板数据在较弱的假设条件（将在 21.4.1 节阐述）下建立因果关系，与之相比，在不含有固定效应的模型（比如混合模型和随机效应模型）中，利用横截面数据或面板数据建立因果关系，则需要较强的假设条件。

在一些研究中，因果关系是清晰明确的，所以随机效应可能是适宜的。在可控实验中，比如源自各种不同数量的肥料用于不同田地的谷物产量，其因果关系是清楚的。在另一些情况下，为了测算相关程度，使用随机效应分析就足够了，而确定因果关系则要采用其他方法做进一步研究。吸烟对肺癌的影响就是一个例子。不过，经济学家却与众不同地偏爱固定效率方法，因为这尽管依赖观测数据，但人们希望测算因果关系。

在实际应用中，固定效应模型拥有几个弱点。对任何时常值回归元，比如性别指示变量的系数进行估计是不可能的，因为它被入列特定个体效应中。而时变回归元的系数是可估计的，只是如果回归元的大部分变异是横截面的而不是随时间变化的，那么这些估计值可能非常不精确。对条件均值进行预测是不可能的，不过，仅有由时变回归元变动而引起的条件均值变化则可以预测。在含有固定效应的非线性模型中，甚至对时变回归元的系数很难加以识别，或者在理论上不可能进行识别。鉴于这些原因，经济学家还是运用随机效应模型，即使因果解释无法得以保证。

21.4.1 *固定效应例子*

考虑计算机使用对工资的影响。几个横截面研究中，最著名的是由克鲁格(Krueger, 1993)与迪纳多和皮施克(DiNardo and Pischke, 1997)研究的那些例子，他们发现，甚至在控制许多决定工资因素诸如教育、年龄、性别、行业以及职业之后，工作中使用计算机与实际较高工资相关联。正如由迪纳多和皮施克(DiNardo and Pischke, 1997)所强调的，如果回归元与误差项相关归因于内生行或省略变量而引起。那么，这不一定蕴含因果关系。

具体地讲，我们假定横截面形式：

$$y_i=\mathbf{x}_i'\boldsymbol{\beta}+\alpha_i+\varepsilon_i$$

其中，y 表示工资自然对数，$\mathbf{x}$ 表示个体特征向量，包括工作中使用计算机的指示变量。而 ε 与 $\mathbf{x}$ 是独立的。一种复杂情况是，添加了不可观测变量 α，假定 α 与工作时使用计算机相关，进而与可观测回归元 $\mathbf{x}$ 相关，尽管 $\mathbf{x}$ 的成分诸如职业与教育而不是使用计算机可部分地控制工作时使用计算机，y 对 $\mathbf{x}$ 的回归产生了省略变量偏倚，从而导致 $\boldsymbol{\beta}$ 非一致估计，因为联合误差($\alpha+\varepsilon$)与 $\mathbf{x}$ 相关。

如果我们假定不可观测变量 α_i 是时常值的，那么围绕该问题，提出一种面板数据方法，α_i 是时常值的。于是，有：

$$y_{it}=\mathbf{x}_{it}'\boldsymbol{\beta}+\alpha_i+\varepsilon_{it}$$

其中，ε 再次与 $\mathbf{x}$ 无关，而 α 与 $\mathbf{x}$ 是相关的。通过进行一阶差分，剔除 α_i（参见 21.2.2 节），使得对 $\boldsymbol{\beta}$ 的一致估计成为可能。就运用计算机例子而言，使用计算机对工资的因果影响可通过个体工资变化与个体进入或离开计算机工作之间的关联加以测算。海斯肯—德纽和施密特（Haisken-DeNew and Schmidt，1999）利用德国面板数据，发现没有影响。

这种固定效应面板方法需要的假设比横截面分析所需要的假设更弱，并以此决定因果关系。其关键性假设是不可观测的 α_i 是时常值的，而不是更一般形式 α_{it}。在计算机使用例子中，假定拥有使用计算机工作的个体倾向是内生的，一旦我们控制了可观测 $\mathbf{x}_{it}$，此倾向 α_i 对工资效应的不可观测成分随时间变化是常值。

一旦我们控制时常值不可观测 α_i 与可观测的 $\mathbf{x}_{it}$，在涉及个体工作是否包括使用计算机的特定时期，基本上被假定成是纯随机的。

随机效应或混合面板方法确实没有类似性质。相反，由于它假定 α 是 iid $[0,\sigma^2]$，进而与 $\mathbf{x}$ 无关，所以它的假设背离了最初关注的 α 与 $\mathbf{x}$ 是相关的内容。倘若 α 实际上与 $\mathbf{x}$ 是相关的，则导致了非一致参数估计，而倘若 α 是时常值的，当 α 与 $\mathbf{x}$ 相关时，则固定效应回归元是一致的。

21.4.2 *条件分析与边际分析*

固定效应估计是条件分析测算一旦控制个体效应 α_i 时 $\mathbf{x}_{it}$ 对 y_{it} 的影响。预测在所用的特殊样本中仅仅对个体而言是可能的，而且甚至如果面板充分长以致能一致地估计出 α_i，那么预测才是可能的。相反，随机效应估计是边际分析或总体平均分析的例子，因为个体效应作为 iid 随机变量可通过积分去掉。随机效应估计量能够用于样本之外。

倘若真实模型是随机效应模型，则是实施条件分析还是实施边际分析将随应用而变化。若分析是针对地区随机样本，则人们使用随机效应，可是倘若人们本质上对样本中特定范围感兴趣，就使用固定效应估计，尽管这会承受有效性损失。

然而，如果真实模型是与回归元相关的特定个体效应的，那么随机效应分析不再有意义，因为随机效应估计量是非一致的。可供选择的其他估计量，比如固定效应估计量与一阶差分估计量却是必需的。由于在微观经济应用中，人们希望决定因果关系，所以才突出后面这一些估计量。

21.4.3 *豪斯曼检验*

如果个体效应是固定的，那么组内估计量 $\hat{\boldsymbol{\beta}}_{\mathrm{W}}$ 是一致的，然而，随机效应估计量 $\tilde{\boldsymbol{\beta}}_{\mathrm{RE}}$是非一致的。此处，$\boldsymbol{\beta}$ 意指时变回归元的系数向量。因此，人们能借助于利用这些估计量豪斯曼检验之间是否存在统计显著差异的，来检验固定效应是否存在。否则，能够使用具有类似性质的任何其他估计量序对，诸如一阶差分与混合 OLS。

大的豪斯曼检验统计量会导致拒绝特定个体效应与回归元不相关的零假设,从而得出结论:固定效应存在。避免利用固定效应模型还是可能的。如果回归元与特定个体效应相关是由省略变量引起的,那么人们可进一步添加回归元,或者是时变的或者是时常值的,然后再次在这种较大模型中执行豪斯曼检验来看看固定效应是否仍是必需的。即使这类相关性持续,但利用工具变量方法估计随机效应模型是可行的(参见 22.4.3~22.4.4 节)。

当 RE 是完全有效时的计算

我们对真实模型做出下述假设来开始,即真实模型是随机效应模型(21.3),其中,α_i iid $[0,\sigma_\alpha^2]$与回归元不相关,并且误差 ε_{it} iid $[0,\sigma_\varepsilon^2]$。

于是,估计量$\tilde{\boldsymbol{\beta}}_{\mathrm{RE}}$是完全有效的,因而由 8.3 节知,豪斯曼检验统计量简化成:

$$\mathrm{H}=(\tilde{\boldsymbol{\beta}}_{1,\mathrm{RE}}-\hat{\boldsymbol{\beta}}_{1,\mathrm{W}})'[\hat{\mathrm{V}}[\hat{\boldsymbol{\beta}}_{1,\mathrm{W}}]-\hat{\mathrm{V}}[\tilde{\boldsymbol{\beta}}_{1,\mathrm{RE}}]]^{-1}(\tilde{\boldsymbol{\beta}}_{1,\mathrm{RE}}-\hat{\boldsymbol{\beta}}_{1,\mathrm{W}})$$

其中,$\boldsymbol{\beta}_1$ 表示对应于时变回归元 $\boldsymbol{\beta}$ 的子成分,因为这个成分可通过组内估计量得到估计。这个检验统计量在零假设下渐近地服从 $\chi^2(\dim[\boldsymbol{\beta}_1])$分布。

豪斯曼(Hausman, 1978)曾经证明,这种检验的渐近等价形式是在辅助 OLS 回归

$$y_{it}-\hat{\lambda}\bar{y}_i=(1-\hat{\lambda})\mu+(\mathbf{x}_{1it}-\hat{\lambda}\,\bar{\mathbf{x}}_{1i})'\boldsymbol{\beta}_1+(\mathbf{x}_{1it}-\bar{\mathbf{x}}_{1i})'\boldsymbol{\gamma}+v_{it} \quad (21.15)$$

中执行沃尔德检验 $\boldsymbol{\gamma}=\mathbf{0}$,其中,$x_{1it}$ 表示时变回归元,而 $\hat{\lambda}$ 已由式(21.11)定义,同时仅使用时变回归元。这个代数结果可作如下解释。特定个体效应模型(21.10)蕴含着,$v_{it}=(1-\hat{\lambda})\alpha_i+(\varepsilon_{it}-\hat{\lambda}\bar{\varepsilon}_i)$。随机效应估计量实际上可通过式(21.15)满足 $\boldsymbol{\gamma}=\mathbf{0}$ 的 OLS 估计来获得[参见式(21.10)]。相反,如果固定效应设定是有效的,α_i 与回归元相关,那么误差 v_{it}将与回归元相关。此种相关性产生了回归元的另外一些函数,比如$(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)$成为式(21.15)中统计显著的变量。

当 RE 不是完全有效时的计算

如果 α_i 或ε_{it}不是 iid 的,这更可能是绝大多数微观经济计量学数据内在性给出的异方差性,那么豪斯曼检验的简单形式就会无效。于是,RE 估计量在零假设下不是完全有效的,因而公式中的表达式 $\hat{\mathrm{V}}[\hat{\boldsymbol{\beta}}_{\mathrm{W}}]-\hat{\mathrm{V}}[\tilde{\boldsymbol{\beta}}_{\mathrm{RE}}]$需要用更一般的$\hat{\mathrm{V}}[\tilde{\boldsymbol{\beta}}_{\mathrm{RE}}-\hat{\boldsymbol{\beta}}_{\mathrm{W}}]$代替(参见 8.3 节)。

对于短面板而言,此方差矩阵能用对不同 i 的自助法重复抽样得到一致估计(参见 21.2.3 节)。因此,面板稳健豪斯曼检验统计量是:

$$\mathrm{H}_{\mathrm{Robust}}=(\tilde{\boldsymbol{\beta}}_{1,\mathrm{RE}}-\hat{\boldsymbol{\beta}}_{1,\mathrm{W}})'[\hat{\mathrm{V}}_{\mathrm{Boot}}[\tilde{\boldsymbol{\beta}}_{1,\mathrm{RE}}-\hat{\boldsymbol{\beta}}_{1,\mathrm{W}}]]^{-1}(\tilde{\boldsymbol{\beta}}_{1,\mathrm{RE}}-\hat{\boldsymbol{\beta}}_{1,\mathrm{W}}) \quad (21.16)$$

其中:

$$\hat{\mathrm{V}}_{\mathrm{Boot}}[\tilde{\boldsymbol{\beta}}_{1,\mathrm{RE}}-\hat{\boldsymbol{\beta}}_{1,\mathrm{W}}]=\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\delta}_b-\bar{\hat{\delta}})(\hat{\delta}_b-\bar{\hat{\delta}})'$$

b 表示 B 次自助复制的第 b 次(参见 21.2.3 节),而 $\hat{\delta}=\tilde{\boldsymbol{\beta}}_{1,\mathrm{RE}}-\hat{\boldsymbol{\beta}}_{1,\mathrm{W}}$。这个检验统计量能用作 $\boldsymbol{\beta}_1$ 的子成分,同时使用可供选择的一些估计量,比如$\tilde{\boldsymbol{\beta}}_{1,\mathrm{POLS}}$代替$\tilde{\boldsymbol{\beta}}_{1,\mathrm{RE}}$,以及$\hat{\boldsymbol{\beta}}_{1,\mathrm{FD}}$代替$\hat{\boldsymbol{\beta}}_{1,\mathrm{W}}$。

否则，伍德里奇(Wooldridge, 2002)建议估计辅助 OLS 回归(21.15)，并利用面板稳健标准误差检验 $\boldsymbol{\gamma}=\mathbf{0}$。如果效应是随机的，虽然不一定使得 α_i 与 ε_{it} 是 iid 的，$v_{it}=(1-\hat{\lambda})\alpha_i+(\varepsilon_{it}-\hat{\lambda}\bar{\varepsilon}_i)$ 还是与回归元不相关，可是 v_{it} 不再是渐近 iid 的，所以需要使用聚集稳健标准误差。如果效应是固定的，那么误差 v_{it} 与回归元相关，导致诸如($\mathbf{x}_{it}-\bar{\mathbf{x}}_i$)回归元的其他函数的显著性。关于豪斯曼检验这种稳健的辅助回归的形式，人们通常假定 v_{it} 是渐近 iid 的，原因在于做出通常的最小化分布假设。然而，当 RE 无效时，人们并不清楚，此种检验实际上是否与豪斯曼检验相符。

豪斯曼检验例子

关于 lnhrs - lnwg 例子，估计值已由表 21.21 给出，利用默认标准误差，对 FE 估计值与 RE 估计值进行比较，得出 $H\simeq(0.168-0.119)^2/(0.019^2-0.014^2)$。从而得出 $H=14>\chi^2_{0.05}(1)=3.84$，所以拒绝随机效应模型。

然而，这种检验是不合适的。因为此例的通常标准误差是非常向下偏倚的(参见 21.3.2 节)，所以统计量 H 被夸大了。而且，此偏倚成为 RE 估计量，在 H_0 条件下，不是完全有效的信号，因此需要使用豪斯曼检验的更一般形式。

由辅助回归(21.15)，得出关于 $\hat{\gamma}$ 的面板稳健 t 统计量为 1.28，从而 $H^*=1.28^2=1.65$，导致在 5%水平上没有拒绝随机效应模型，即使工资弹性估计值相差 0.049，但该估计非常不精确，其差并不是统计显著的。注意到，如果使用关于 $\hat{\gamma}$ 的非稳健 t 统计量，那么 $t^2=13.69$，接近于前面不正确的豪斯曼检验统计量。

21.4.4 *较丰富的随机效应模型*

随机效应模型设定随机效应 α_i 是回归元的独立分布。较丰富的模型在思想上更接近于固定效应模型，它放松了这一假设。

德拉克允许面板模型(21.3)中的个体效应可由回归元的时间平均来决定，因而 $\alpha_i=\bar{\mathbf{x}}_i'\boldsymbol{\pi}+w_i$，其中，$w_i$ 是 iid 的，于是，在这种扩展模型中，$\boldsymbol{\beta}$ 与 $\boldsymbol{\pi}$ 的有效 GLS 估计会得出 $\boldsymbol{\beta}$ 估计量，该估计量等于模型(21.3)的固定效应估计量。通过比较发现，错误设定 iid 随机效应的模型(21.3)中的 $\boldsymbol{\beta}$ 随机效应估计量将是非一致的。

张伯伦(Chamberlain, 1982, 1984)考察了随机效应的更为丰富的模型，满足 $\alpha_i=\mathbf{x}_{1i}'\boldsymbol{\pi}_1+\cdots+\mathbf{x}_{Ti}'\boldsymbol{\pi}_T+w_i$，即回归元的加权和。他提出通过最小距离方法加以估计(详细内容参见 22.2.7 节)，导致了等于固定效应估计量的 $\boldsymbol{\beta}$ 的估计量。

更一般地，24.6 节的混合线性模型与分层线性模型允许含有随机截距，而且可包含随机斜率参数的相当一般模型。面板数据的贝叶斯分析也可使用这种框架。详细内容参见 22.8 节。

在线性模型中，若不可观测个体效应与回归元相关，则运用固定效应方法。在更复杂模型中，例如非线性模型，固定效应模型并不总是可估计的，但较丰富随机效应模型却提供了可供选择的方法。

21.5 混合模型

混合横截面时间序列模型(**pooled cross-section time-series model**)或常系数模型(**constant-coefficients model**)是：

$$y_{it}=\alpha+\mathbf{x}_{it}'\boldsymbol{\beta}+u_{it} \tag{21.17}$$

在统计学文献中,此模型称为总体平均模型(**population-averaged model**),因为以个体效应为条件的 y_{it} 的显式模型不存在。相反,任何个体效应都可以用隐性方式加以平均去掉。随机效应模型是下述特殊情况:给定 i 时,误差 u_{it} 关于不同时间是等相关的(参见 21.2.1 节)。

一旦假定没有固定效应,统计推断的主要复杂情况是,模型普通最小二乘法估计量的分布会随对 u_{it} 假定的分布而变化。在短面板中,能利用式(21.13)获得面板稳健标准误差。

然而,我们此处关注利用各种不同设定的 GLS 估计,包括等相关性,因为文献提出了关于不同时间与个体的 u_{it} 的协方差结构。

虽然我们关注式(21.17)即不含有特定个体固定效应的混合 GLS 估计,但本节方法通常能用于 21.2.3 节变换模型(21.12)的混合 GLS 估计。

21.5.1 *混合 OLS, FGLS 以及 WLS 估计量*

利用矩阵记号表述极为方便。对于给定个体,将不同时间观测值组合起来定义:

$$\mathbf{y}_i=\mathbf{W}_i\boldsymbol{\delta}+\mathbf{u}_i \tag{21.18}$$

其中,$\boldsymbol{\delta}=[\alpha\ \boldsymbol{\beta}']'$ 表示 $(K+1)\times1$ 维参数向量,$\mathbf{y}_i$ 与 $\mathbf{u}_i$ 分别表示第 t 个元素为 y_{it} 与 u_{it} 的 $T\times1$ 维向量,$\mathbf{W}_i$ 表示 $T\times(K+1)$ 阶矩阵,其第 t 行表示 $\mathbf{w}_{it}'=[1\quad \mathbf{x}_{it}]'$。若对所有个体进行叠放,得到:

$$\mathbf{y}=\mathbf{W}\boldsymbol{\delta}+\mathbf{u}_i \tag{21.19}$$

其中,$\mathbf{y}$ 与 $\mathbf{u}$ 表示 $NT\times1$ 维向量,例如 $\mathbf{y}=[\mathbf{y}_1'\cdots\mathbf{y}_N']$,而 $\mathbf{W}$ 表示 $NT\times(K+1)$ 阶回归元矩阵,其第 1 列为单位向量。我们假定 $\mathrm{E}[\mathbf{u}|\mathbf{W}]=\mathbf{0}$,所以误差是严格外生的,同时定义 $\boldsymbol{\Omega}=\mathrm{E}[\mathbf{uu}'|\mathbf{W}]$。

这个模型存在几种可能的最小二乘法估计量,已概述在表 21.5 中。

第一,混合 OLS 是一致的且渐近正态的。然而,在面板背景下,不可能有 $\boldsymbol{\Omega}=\sigma^2\mathbf{I}_{NT}$,所以除了某些特殊情况,诸如当所有回归元都是时常值时,OLS 是无效的。更为重要的是,不应运用 $\sigma^2(\mathbf{W}'\mathbf{W})^{-1}$ 通常方差估计,而应运用式(21.13)所需的那种面板稳健估计。

第二,混合可行 GLS(**pooled feasible GLS**, 记为 PFGLS)是一致的且完全有效的,如果 $\boldsymbol{\Omega}$ 被正确地设定,同时 $\hat{\boldsymbol{\Omega}}$ 关于 $\boldsymbol{\Omega}$ 是一致的。面板文献对 u_{it} 结构提出了某种非常大范围的要求,从而对 $\boldsymbol{\Omega}$ 结构也施加某种要求,这些已经被并入分别由 21.5.2 节与 21.5.3 节给出的关于短面板与长面板的回归软件包之中。

第三,混合加权 LS(**pooled weighted LS**, 记为 PWLS)估计量防止了对 $\boldsymbol{\Omega}$ 的错误设定。对误差方差矩阵 $\boldsymbol{\Omega}$,该估计量假定有一个实用矩阵 $\boldsymbol{\Sigma}$,然后进行推断,甚至当 $\boldsymbol{\Sigma}\neq\boldsymbol{\Omega}$ 时,该推断仍是有效的。普通最小二乘法是满足 $\boldsymbol{\Sigma}=\sigma^2\mathbf{I}_{NT}$ 的例子,而对 $\boldsymbol{\Sigma}$ 的其他选择可能提高有效性。

混合 OLS 估计量的方差矩阵估计需要 $\hat{\boldsymbol{\Omega}}$,使得 $(NT)^{-1}\mathbf{W}'\hat{\boldsymbol{\Omega}}\mathbf{W}$ 一致地估计

$(NT)^{-1}\mathbf{W}'\mathbf{\Omega W}$。

对于短面板数据来说，这可通过直接应用 21.2.3 节的结果执行。混合 WLS 估计量的方差矩阵估计需要 $\hat{\mathbf{\Omega}}$，使得 $(NT)^{-1}\mathbf{W}'\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{\Omega}}\hat{\mathbf{\Sigma}}^{-1}\mathbf{W}$ 一致地估计 $(NT)^{-1}\mathbf{W}'\mathbf{\Sigma}^{-1}\mathbf{\Omega\Sigma}^{-1}\mathbf{W}$。由式(21.13)给出的关于 OLS 面板稳健估计，借助于替换 $\mathbf{W}'\mathbf{\Sigma}^{-1}\mathbf{\Omega\Sigma}^{-1}\mathbf{W}$，或者等价地借助于数量 $\sum_i \mathbf{W}_i'\mathbf{\Sigma}_i^{-1}\mathrm{E}[\mathbf{u}_i\mathbf{u}_i'\mid\mathbf{W}_i]\mathbf{\Sigma}_i^{-1}\mathbf{W}_i$ 替换给定不同 i 时具有独立性的 $\sum_i \mathbf{W}_i'\hat{\mathbf{\Sigma}}_i^{-1}\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'\mathbf{\Sigma}_i^{-1}\mathbf{W}_i$，而适应混合 WLS，其中，$\hat{\mathbf{u}}_i=\mathbf{y}_i-\mathbf{W}_i\hat{\boldsymbol{\delta}}$。否则，使用面板自助法。

21.5.2 短面板的误差方差矩阵

在短面板中，存在几个时期但众多个体，通常是人或厂商。假定误差对不同个体而言是独立的，因此 $\mathrm{Cov}[u_{it},u_{js}]=0, i\neq j$。在此情况下，重新运用求和记号非常方便。例如，由表 21.5 给出的 PFGLS 估计量变成：

$$\hat{\beta}_{\mathrm{PFGLS}}=\left[\sum_{i=1}^{N}\mathbf{W}_i'\hat{\mathbf{\Omega}}_i^{-1}\mathbf{W}_i\right]^{-1}\sum_{i=1}^{N}\mathbf{W}_i'\hat{\mathbf{\Omega}}_i^{-1}\mathbf{y}_i \tag{21.20}$$

其中，$\hat{\mathbf{\Omega}}_i$ 关于

$$\mathbf{\Omega}_i=\mathrm{E}[\mathbf{u}_i\mathbf{u}_i'\mid\mathbf{W}_i] \tag{21.21}$$

是一致的，而 $\mathbf{\Omega}_i$ 是非对角的，因为误差对给定个体而言可能对不同时间是相关的。注意到，$\hat{\mathbf{\Omega}}_i$ 必须来自对 $\mathbf{\Omega}_i$ 设定模型的估计，而且不能使用 $\hat{\mathbf{\Omega}}_i=\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'$ [参见式(5.88)之后的有关讨论]。

表 21.5　混合最小二乘法估计量及其渐近方差

估计量	公式[a]	方差矩阵[b]
混合 OLS：$\hat{\boldsymbol{\delta}}_{\mathrm{POLS}}$	$(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{y}$	$(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\hat{\mathbf{\Omega}}\mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}$
混合 FGLS：$\hat{\boldsymbol{\delta}}_{\mathrm{PFGLS}}$	$(\mathbf{W}'\hat{\mathbf{\Omega}}^{-1}\mathbf{W})^{-1}\mathbf{W}'\hat{\mathbf{\Omega}}^{-1}\mathbf{y}$	$(\mathbf{W}'\hat{\mathbf{\Omega}}^{-1}\mathbf{W})^{-1}$
混合 WLS：$\hat{\boldsymbol{\delta}}_{\mathrm{PWLS}}$	$(\mathbf{W}'\hat{\mathbf{\Sigma}}^{-1}\mathbf{W})^{-1}\mathbf{W}'\mathbf{\Sigma}^{-1}\mathbf{y}$	$(\mathbf{W}'\hat{\mathbf{\Sigma}}^{-1}\mathbf{W})^{-1}\mathbf{W}'\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{\Omega}}\hat{\mathbf{\Sigma}}^{-1}\mathbf{W}(\mathbf{W}'\hat{\mathbf{\Sigma}}^{-1}\mathbf{W})^{-1}$

[a] 公式是式(21.19)定义的模型 $\mathbf{y}=\mathbf{W}\boldsymbol{\delta}+\mathbf{u}$，而误差矩阵为 $\mathbf{\Omega}$。

[b] 为了计算 POLS 与 PWLS 的方差矩阵，参见正文内容，在那些情况下，$\hat{\mathbf{\Omega}}$ 关于 $\hat{\mathbf{\Omega}}$ 不必是一致的。就混合 AFGLS 而言，假定 $\hat{\mathbf{\Omega}}$ 关于 $\mathbf{\Omega}$ 是一致的。

等相关误差

最广泛使用的误差结构是 21.2.1 节曾阐述的随机效应模型。于是，由式(21.6)知，$\mathbf{\Omega}_i$ 具有共同对角元素 $\sigma_\alpha^2+\sigma_\varepsilon^2$，以及共同非对角元素 σ_α^2。等价地，因为 $\mathbf{\Omega}_i$ 具有共同对角元素 σ^2 与共同性对角元素 $\rho\sigma^2$，故此误差是等相关的。实施 AFGLS 仅仅需要估计 σ_α^2 与 σ_ε^2，或者估计 σ^2 与 ρ(参见 21.2.2 节与 21.7 节)。

ARMA 误差

一种可供选择的误差结构是假定 ARMA 误差模型。例如，AR(1)误差模型设定 $u_{it}=\rho u_{i,t-1}+\varepsilon_{it}$，其中，$\varepsilon_{it}$ 是 iid 的。于是，$\mathrm{Cov}[u_{it},u_{is}]=\rho^{|t-s|}\sigma^2$。在此情况下，当误差之间的时期数增大时，误差之间的协方差就下降了。RE 模型与 AR(1)误差模型的比较将由 21.5.4 节给出。

巴尔塔吉和李(Baltagi and Li, 1991)为了考察含有 AR(1)误差的随机效应模型,将两种误差模型结合起来。这很容易推广到 AR(p)情况,而且随机效应模型中关于移动平均的一些方法以及 ARMA 误差目前也得到了发展。巴尔塔吉(Baltagi, 2001,第 5 章)给出了一个概括。

带有非结构化自相关的同方差误差

对于 FGLS 估计来说,无结构化自相关的同方差误差,做出 $T\times T$ 阶矩阵 $\mathbf{\Omega}_i$ 对不同 i 而言都是常值的假设,则短面板数据实际上不需要施加诸如由 RE 模型或 AR(1)误差模型所施加的更多结构。于是,需要估计的参数"仅有"$T(T+1)/2$ 个。于是,$\mathbf{\Omega}_i$ 的一致估计是 $\hat{\mathbf{\Omega}}_i$,它的第(t,s)个元素为 $\hat{\sigma}_{ts}=N^{-1}\sum_{i=1}^{N}\hat{u}_{it}\hat{u}_{is}$。前面的模型同样假定了同方差性,只是对 $\mathbf{\Omega}_i$ 设置了其他结构。

稳健推断

前面的所有设定均假定,误差协方差对不同个体而言都是相同的,这样就剔除了异方差型。倘若面板是短的,人们仍然能使用上面的约束误差方差矩阵模型作为混合 WLS 估计的基础,但在另一方面获得如同表 21.5 之后讨论的稳健标准误差。否则,使用第 22 章阐述的较丰富的混合模型进行估计。

第 21 章至第 23 章自始至终保持对不同 i 具有独立性的假设,尽管倘若对相关性施加结构,甚至对小 T 来说,可放松上述假设。一个实例是关于空间相关的明显模型,这里的空间相关涉及地区面板数据,比如州或区域,当个体之间的自然距离增大时,其相关性会下降。

21.5.3 *长面板误差方差矩阵*

在长面板中存在许多时期,但具有相对很少的个体。如果个体观测单位是少数几个地区之一,诸如州或地区或厂商,那么这类数据便出现在微观经济计量学分析之中,但为了将推断建立在 $T\to\infty$假设的基础上,这些在足够多时期上都是可观测到的。

对于给定个体来说,其不同时期的相关性可利用误差的 ARMA 模型来引进,ARMA 模型的参数允许当目前 N 为固定且 $T\to\infty$时,对不同个体而言是不同的。例如,考察具有 $u_{it}=\rho_i u_{i,t-1}+\varepsilon_{it}$ 的 AR(1)误差,其中 $\varepsilon_{it}\sim[0,\sigma_i^2]$表示异方差的,并且 ρ_i 对不同个体而言也是不一样的。分别将 y_{it} 对 $\mathbf{w}_{it}$ 进行回归,就每一个个体而言,因为 $T\to\infty$,利用 T 个时期的 AR(1)误差都会得出一致估计值 $\hat{\rho}_i$ 与 $\hat{\sigma}_i^2$。从而,当有 NT 个观测值时,这用于对 $\boldsymbol{\delta}$ 的可行 GLS 估计。详细内容,参见克门塔(Kmenta, 1986)。这个模型既允许个体具有异方差性,又允许对给定个体来说具有时期相关性。佩萨兰(Pesaran,2004)提出了借助于 GLS 进行估计的相当多的更丰富模型。

对于长面板,引入不同个体相关性是可行的,因此对于 $i\neq j$,$\mathrm{Cov}[u_{it},u_{is}]\neq 0$,由于 N 是固定的,且渐近结果依赖于 $T\to\infty$。特别地,人们能像前面一样满足不同个体之间独立性假设执行混合 GLS 估计,但要利用 6.4.4 节曾简要提及的纽韦和韦斯特(Newy and West, 1987b)方法计算标准误差,倘若序列相依性充分快速衰减,这样做,允许出现任意横截面相依性与序列相依性。详细内容,参见阿雷拉

诺(Arellano, 2003,第 19 页)。

21.5.4 自相关误差的影响

面板数据回归模型具有下述误差,对给定个体而言经常具有不同时期的相关性。若没有固定效应,则混合 OLS 回归会得出一致参数估计值。不过,当忽略自相关时,误差相关能导致混合 OLS 的标准误差出现大的偏倚,而当面板长度增加时,其有效性提高相对很小。

就单个个体(因此,$N=1$)具有等相关而言,对基于 T 个观测值的对 y 均值进行估计来说,分析起来特别简单。于是,$y_t=\beta+u_t$,而 OLS 估计量是样本均值,所以 $\hat{\beta}=\bar{y}=T^{-1}\sum_t y_t$。该 OLS 估计量具有真实方差 $\mathrm{V}[\hat{\beta}]=\mathrm{V}[\bar{y}]=T^{-2}\sum_t\sum_s \mathrm{Cov}[u_t,u_s]$。若假定等相关,则此双和式具有等于 σ^2 的 T 个方差,以及等于 $\rho\sigma^2$ 的 $T(T-1)$ 个协方差。因此,$\mathrm{V}[\bar{y}]=T^{-1}\sigma^2(1+(T-1)\rho)$。因而,$\mathrm{V}[\bar{y}]=T^{-1}\sigma^2$ 的 iid 结果需要借助乘以$(1+\rho(T-1))$加以扩大而得到修改,特别地,当 $\rho\to 1$ 时,$\mathrm{V}[\bar{y}]$趋向于 σ^2。

对于各种 T 与 ρ 值,表 21.6 给出关于 $\bar{y}$ 方差的相关影响,这里,为了简单起见,我们正规化 $\sigma^2=1$,当 ρ 增大时,估计准确性大大下降,并在给定第一列独立性假设(为了简单起见,假定 σ^2 是已知的)条件下,$\mathrm{V}[\bar{y}]$估计值远远低估了真实方差。此外,对于 $\rho>0$,因时期数增大所获得的准确性提高远不及因独立数据引起的准确性提高,倍增时期数会使估计量方差减半。例如,若 $\rho=0.4$,则 5 个时期数的估计量方差只是 1 个时期数估计量方差的 0.52 倍,不过,而不是独立数据时的 0.2 倍这一更小情况。进一步地,若从 5 个时期数到 10 个时期数增加 1 倍,则会得到估计量方差从 0.52 到 0.46,只是出现很小缩减。

表 21.6 含有等相关误差的混合 OLS 估计量方差[a]

T	$\rho=0.0$	$\rho=0.2$	$\rho=0.4$	$\rho=0.6$	$\rho=0.8$	$\rho=1.0$
1	1.00	1.00	1.00	1.00	1.00	1.00
2	0.50	0.60	0.70	0.80	0.90	1.00
5	0.20	0.36	0.52	0.68	0.84	1.00
10	0.10	0.28	0.46	0.64	0.82	1.00

[a] 当等相关误差的相关 ρ 增大时,给出了混合 OLS 估计量的方差,对于具有误差方差被正规化为 1 的唯一截距模型来说,尽管是同方差的,但假定误差是相关的。

对于更一般的具有等相关误差且回归元为时常值的平衡面板回归来说,这个结果成立,其中,OLS 估计量的真实方差是假定独立误差情况的$(1+\rho(T-1))$倍[参见克勒克(Kloek,1981)]。在实际应用中,还会包括时变回归元,而且很明显,非常难以获得解释结果。对于含有截距与单个时变回归元的回归来说,斯科特和霍尔特(Scott and Holt, 1982)证明,斜率系数的方差扩大了$(1+\hat{\rho}_x\rho(T-1))$倍,其中,$\hat{\rho}_x$ 被认为是特定个体 x 的自相关的估计值,就面板数据而言,$\hat{\rho}_x$ 往往很高,因此仍然存在明显的扩大。这些结果同样可应用于聚集数据的其他形式,更详细的内容将在 24.5.2 节论述。

前面分析均假定等相关误差,即 RE 模型的性质。相反,若误差是 AR(1)的,

则增加面板长度会有很大好处。于是，$\mathrm{Cov}[u_t, u_s] = \rho^{|t-s|}\sigma^2$，因而，$V[\bar{y}] = T^{-2}\sigma^2[T + 2\sum_{s=1}^{T-1}(T-s)\rho^s]$。例如，当 $\rho=0.8$ 时，对于 $T=5$，$V[\bar{y}]=0.72\sigma^2$，而对于 $T=10$，$V[\bar{y}]=0.54\sigma^2$，这均小于满足 $\rho=0.8$ 等相关的表 21.6 中相应 $0.84\sigma^2$ 与 $0.82\sigma^2$ 的值，但仍远远大于 $\rho=0.0$ 的相应 $0.2\sigma^2$ 与 $0.1\sigma^2$ 的值。

微观经济计量学家倾向于 RE 模型或短面板的等相关误差模型，如同第 24 章对聚集数据派生文献所阐述的。例如，考察许多家庭里不同的兄弟姐妹。于是，自然假定同一个家庭中不同胞亲不可观测因素相关性，对于不同兄弟姐妹均是一样的。例如，老大与老二之间的相关性等于老大与老三之间的相关性。相反，那些利用长面板数据经常具有时间序列背景，并很自然地假定相关性随时间而下降，得到诸如 AR(1)误差的模型。

实际上，决定时间序列相关的哪一种模型更为合理，这要依赖数据而定。微观经济计量学应用所使用的短面板会得出混合 OLS 残差自相关，在性质上类似于由表 21.3 给出的那些情况。这些比较接近于 RE 模型，而不是 AR(1)模型，尽管 ARMA(1, 1)可能做得很好。含有 AR(1)误差的 RE 模型仍然更好一些。在所有情况下，误差相关性会引起信息损失，通常 OLS 标准误差低估了真实标准误差。对于短面板数据，人们能将推断建立在面板稳健标准误差上(参见 21.2.3 节)，而不需要设定误差相关模型。

21.5.5 *小时与工资混合GLS例子*

表 21.7 给出了 lnhrs 对 lnwg 回归的模型 $y_{it} = \alpha_i + \beta x_{it} + u_{it}$ 的一系列混合 GLS 估计值，以及与之有关的默认表误差与稳健标准误差。

表 21.7 小时与工资：混合 OLS 与 GLS 估计值[a]

估计量	POLS	PFGLS		
误差相关	没有	等相关	AR1	一般的
α	7.442	7.346	7.440	7.426
β	0.083	0.120	0.084	0.091
稳健 se	(0.029)	(0.052)	(0.037)	(0.050)
自助 se	[0.032]	[0.060]	[0.050]	[—]
默认 se	{0.009}	{0.014}	{0.012}	{0.014}

[a] 对于短面板，若假定对于不同 i 具有独立性且同分布的，同时没有固定效应，lnhrs 对 lnwg 混合 OLS 与 GLS 线性面板回归。混合 GLS 估计量假定等相关的或随机效应误差(equi)、AR(1)误差(AR(1))或者没有相关性结构(一般情况)。斜率系数的标准误差若是面板稳健的，用圆括号表示；面板自助法标准误差用方括号表示；而假定 iid 误差的通常默认估计则用大括号表示。

所有内容均假定误差 u_{it} 对不同 i 是独立的，且对不同 i 是同分布的，然后对不同 t 做出 u_{it} 相关性的各种不同假设。

表 21.7 的第 1 列，即混合 OLS 估计量，重新列出表 21.2 的第 1 列内容：混合 GLS 估计假定等相关，由表 21.7 第 2 列给出。这些均与表 21.2 的 RE - GLS 列相吻合，因为随机效应模型蕴含着等相关误差[参见式(21.6)]。

混合 GLS 估计假定 AR(1)误差，所以 $u_{it} = \rho u_{i,t-1} + \varepsilon_{it}$，其中，$\varepsilon_{it}$ 是 iid 的，这由表 21.7 中第 3 列给出。其斜率与系数估计值比较接近于混合 OLS 估计值。

除同方差性之外，对误差相关没有施加结构的混合 GLS 估计值，均已由表 21.7 第 4 列给出，因此 $\mathrm{Cov}[u_{it},u_{is}]=\sigma_{ts}$。于是，给定很小 T，通过 $\hat{\sigma}_{ts}=N^{-1}\sum_{i=1}^{N}\hat{u}_{it}\hat{u}_{is}$，可一致估计出 σ_{ts}，对于所有的 t 与 s。这再次接近于混合 OLS 估计值。

由表 21.7 清楚知道，应该使用面板稳健标准误差，而不是使用默认标准误差，这里假定同方差性，并且正确地设定序列相关模型。

21.6 固定效应模型

固定效应模型(**fixed effects models**)设定：

$$y_{it}=\alpha_i+\mathbf{x}_{it}'\boldsymbol{\beta}+\varepsilon_{it} \tag{21.22}$$

其中，特定个体效应 $\alpha_1,\cdots,\alpha_N$ 测量了不可观测异质性，异质性可能与回归元 $\mathbf{x}_{it}$ 相关，$\boldsymbol{\beta}$ 表示 $K\times1$ 维向量，而且以误差服从 iid $[0,\sigma^2]$开始讨论。

进行估计的一个挑战是，存在 N 个特定个体效应，而当 $N\to\infty$时，这 N 个特定个体效应会增加。考虑到应用目的，我们对 K 个斜率参数 $\boldsymbol{\beta}$ 最感兴趣，$\boldsymbol{\beta}$ 给出回归元变化时的边际效应，因为$\partial\mathrm{E}[y_{it}]/\partial\mathbf{x}_{it}=\boldsymbol{\beta}$。$N$ 个参数 $\alpha_1,\cdots,\alpha_N$ 是冗余参数或非主要参数[1](**incidental parameters**)，但它们不是人们内在关注的内容。不过，它们的存在会潜在地阻碍对参数 $\boldsymbol{\beta}$ 进行估计，而 $\boldsymbol{\beta}$ 是关注内容。

值得注意的是，尽管存在这些冗余参数，但线性模型存在几种一致估计 $\boldsymbol{\beta}$ 的方法。这些方法包括：(1) 组内模型(21.8)的 OLS；(2) 对 N 个固定效应中的每一个都具有指示变量的模型(21.2)进行直接 OLS 估计；(3) 组内模型(21.8)的 GLS；(4) 以个体均值 $\bar{y}_i$ 为条件的 ML 估计，$i=1,\cdots,N$；(5)一阶差分模型(21.9)的 OLS。

前两种方法总是得出 $\boldsymbol{\beta}$ 的相同估计量。另外，如果式(21.22)中的 ε_{it} 是 iid 的，并且 $\varepsilon_{it}\sim\mathcal{N}[0,\sigma^2]$，那么第三种方法与第四种方法也将是一样的。对于 $T>2$，最后一种方法则不同于其他方法。在非线性模型里，这种等价性经常不成立，这将在第 23 章讨论。

下一节给出组内估计量的基本结果。当回归元不再是强外生时，21.6.2 节阐述一阶差分估计量，该估计量广泛用于第 22 章。而其他估计量将由 21.6 节的其余部分加以阐述，一些读者或许愿意略过它们。

21.6.1 *组内或固定效应估计量*

组内模型可通过从最初模型中减去时间平均模型 $\bar{y}_i=\alpha_i+\bar{\mathbf{x}}_i'\boldsymbol{\beta}+\bar{\varepsilon}_i$而得到，于是：

$$y_{it}-\bar{y}_i=(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)'\boldsymbol{\beta}+(\varepsilon_{it}-\bar{\varepsilon}_i) \tag{21.23}$$

因此，固定效应 α_i 被剔除，如果对于所有 t，$\mathbf{x}_{it}=\mathbf{x}_i$，由于 $\mathbf{x}_{it}-\bar{\mathbf{x}}_i=\mathbf{0}$，所以时常值回归元也被剔除了。

---

[1] 又称为偶发参数。——译者注

利用 OLS 估计，得到组内估计量或固定效应估计量 $\hat{\boldsymbol{\beta}}_W$，其中：

$$\hat{\boldsymbol{\beta}}_W=\Big[\sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)'\Big]^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)(y_{it}-\bar{y}_i) \tag{21.24}$$

于是，个体固定效应 α_i 能通过：

$$\hat{\alpha}_i=\bar{y}_i-\bar{\mathbf{x}}_i'\hat{\boldsymbol{\beta}}_W,\quad i=1,\cdots,N \tag{21.25}$$

加以估计，估计值 $\hat{\alpha}_i$ 关于 α_i 是无偏的，倘若 $N\to\infty$，它就是一致的，因为 $\hat{\alpha}_i$ 对 T 个观测值进行平均。在短面板中，估计值 $\hat{\alpha}_i$ 是非一致的，但 $\hat{\boldsymbol{\beta}}_W$ 关于 $\boldsymbol{\beta}$ 却是一致的。可以认为，α_i 是冗余参数或辅助参数，幸运的是，为了获得更重要参数 $\boldsymbol{\beta}$ 的一致估计值，而不要求一致地估计 α_i。

组内估计量的一致性

当 $\text{plim}(NT)^{-1}\sum_i\sum_t(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)(\varepsilon_{it}-\bar{\varepsilon}_i)=\mathbf{0}$ 时，$\boldsymbol{\beta}$ 的组内估计量是一致的。倘若要么 $N\to\infty$ 要么 $T\to\infty$，并且：

$$\mathrm{E}[\varepsilon_{it}-\bar{\varepsilon}_i\,|\,\mathbf{x}_{it}-\bar{\mathbf{x}}_i]=0 \tag{21.26}$$

则应是这种情况。由于平均值 $\bar{\mathbf{x}}_i=T^{-1}\sum_t\mathbf{x}_{it}$ 与 $\bar{\varepsilon}_i$ 都存在，所以此条件：$\mathrm{E}[\varepsilon_{it}\,|\,\mathbf{x}_{it}]=0$ 是较强的。式(21.26)的充分条件是强外生性条件 $\mathrm{E}[\varepsilon_{it}\,|\,\mathbf{x}_{i1},\cdots,\mathbf{x}_{it}]=0$。这就排除了含有滞后内生变量作为回归元的组内估计(参见 22.5 节)。

组内估计量的渐近分布

由于对于给定 i 时组内模型(21.8)的误差$(\varepsilon_{it}-\bar{\varepsilon}_i)$关于 t 是相关的，所以 $\hat{\boldsymbol{\beta}}_W$ 分布潜在表现出复杂性。下面将证明这一结论，并应用通常的 OLS 结果。在强假设下，即 ε_{it} 是 iid 的，有：

$$\mathrm{V}[\hat{\boldsymbol{\beta}}_W]=\sigma_\varepsilon^2\Big[\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{\mathbf{x}}_{it}\ddot{\mathbf{x}}_{it}'\Big]^{-1} \tag{21.27}$$

其中，$\ddot{\mathbf{x}}_{it}=\mathbf{x}_{it}-\bar{\mathbf{x}}_i$。$\sigma_\varepsilon^2$ 的一致且无偏估计是 $\hat{\sigma}_\varepsilon^2=[N(T-1)-K]^{-1}\sum_i\sum_t\hat{\varepsilon}_{it}^2$，其中，自由度等于样本量 NT 减去模型参数个数 T，再减去 N 个个体效应。注意，若利用标准最小二乘法软件包估计回归(21.23)，则需要通过 $[N(T-1)-K]^{-1}[NT-K]$ 增大报告方差。

对于短面板，式(21.13)得出渐近方差的稳健估计：

$$\mathrm{V}[\hat{\boldsymbol{\beta}}_W]=\Big[\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{\mathbf{x}}_{it}\ddot{\mathbf{x}}_{it}'\Big]^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\ddot{\mathbf{x}}_{it}\ddot{\mathbf{x}}_{is}'\hat{\ddot{\varepsilon}}_{it}\hat{\ddot{\varepsilon}}_{is}\Big[\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{\mathbf{x}}_{it}\ddot{\mathbf{x}}_{it}'\Big]^{-1} \tag{21.28}$$

其中，$\ddot{\varepsilon}_{it}=\varepsilon_{it}-\bar{\varepsilon}_i$。这种深受人们偏爱的估计允许 ε_{it} 的任意自相关以及任何异方差性。

推导组内估计量方差

现在，利用矩阵代数推导式(21.27)给出的组内估计量方差估计。我们以第 i 个观测值的模型

$$y_{it}=\alpha_i+\mathbf{x}_{it}'\boldsymbol{\beta}+\varepsilon_{it}$$

开始，其中，$\mathbf{x}_{it}$ 与 $\boldsymbol{\beta}$ 都表示 $K\times1$ 维向量。对于第 i 个个体，将所有 T 个观测值加以叠放，则有：

$$\begin{bmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \alpha_i + \begin{bmatrix} \mathbf{x}_{i1}' \\ \vdots \\ \mathbf{x}_{iT}' \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{iT} \end{bmatrix}, \quad i=1,\cdots,N$$

或

$$\mathbf{y}_i = \mathbf{e}\alpha_i + \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \quad i=1,\cdots,N \tag{21.29}$$

其中，$\mathbf{e}=(1,1,\cdots,1)'$表示所有元素为 1 的 $T\times1$ 维单位向量，$\mathbf{X}_i$ 表示 $T\times K$ 阶矩阵，而 $\mathbf{y}_i$ 与 $\boldsymbol{\varepsilon}_i$ 均表示 $T\times1$ 维向量。

为了将模型(21.29)变换成组内模型，就要减去特定个体均值，引入 $T\times T$ 阶矩阵：

$$\mathbf{Q} = \mathbf{I}_T - T^{-1}\mathbf{e}\mathbf{e}' \tag{21.30}$$

利用矩阵 $\mathbf{Q}$ 左乘，则得到离差，因为：

$$\mathbf{Q}\mathbf{W}_i = \mathbf{W}_i - \mathbf{e}\bar{\mathbf{w}}_i' \tag{21.31}$$

其中，$\mathbf{W}_i$ 表示 $T\times M$ 矩阵，其第 t 行为 $\mathbf{w}_{it}'$，而 $\bar{\mathbf{w}}_i = T^{-1}\sum_{t=1}^{T}\mathbf{w}_{it}$ 表示 $m\times1$ 维平均向量。利用 $\mathbf{e}'\mathbf{W}_i = T\bar{\mathbf{w}}_i'$，可获得结果(21.31)。如果利用 $\mathbf{e}'\mathbf{e}=T$ 且 $\mathbf{Q}\mathbf{e}=\mathbf{0}$，那么 $\mathbf{Q}\mathbf{Q}'=\mathbf{Q}$，所以 $\mathbf{Q}$ 是幂等的。

通过 $\mathbf{Q}$ 左乘第 i 个个体的固定效应模型(21.29)，利用 $\mathbf{Q}\mathbf{e}=\mathbf{0}$，得出：

$$\mathbf{Q}\mathbf{y}_i = \mathbf{Q}\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Q}\boldsymbol{\varepsilon}_i, \quad i=1,\cdots,N \tag{21.32}$$

这是组内模型(21.23)，因而通过 $\mathbf{Q}$ 左乘，得到组内估计量。当假设对于不同 i 具有独立性时，对式(21.32)进行 OLS 估计，从而得到具有方差矩阵的 $\hat{\boldsymbol{\beta}}_W$，等于：

$$\mathrm{V}[\hat{\boldsymbol{\beta}}_W] = \Big[\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{Q}'\mathbf{Q}\mathbf{X}_i\Big]^{-1}\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{Q}'\mathrm{V}[\mathbf{Q}\boldsymbol{\varepsilon}_i\,|\,\mathbf{X}_i]\mathbf{Q}\mathbf{X}_i\Big[\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{Q}'\mathbf{Q}\mathbf{X}_i\Big]^{-1} \tag{21.33}$$

若以强假设：ε_{it}是 iid $[0,\sigma_\varepsilon^2]$开始讨论，则 $\boldsymbol{\varepsilon}_i$ 是 iid $[\mathbf{0},\sigma_\varepsilon^2\mathbf{I}]$。于是，$T\times1$ 维误差 $\mathbf{Q}\boldsymbol{\varepsilon}_i$ 关于不同 i 是独立的，其均值为 0，且方差 $\mathrm{V}[\mathbf{Q}\boldsymbol{\varepsilon}_i]=\mathbf{Q}\mathrm{V}[\boldsymbol{\varepsilon}_i]\mathbf{Q}'=\sigma_\varepsilon^2\mathbf{Q}\mathbf{Q}'=\sigma_\varepsilon^2\mathbf{Q}$ 。从而：

$$\begin{aligned}\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{Q}'\mathrm{V}[\mathbf{Q}\boldsymbol{\varepsilon}_i\,|\,\mathbf{X}_i]\mathbf{Q}\mathbf{X}_i &= \sum_{i=1}^{N}\mathbf{X}_i'\mathbf{Q}'\sigma_\varepsilon^2\mathbf{Q}\mathbf{Q}\mathbf{X}_i \\ &= \sigma_\varepsilon^2\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{Q}'\mathbf{Q}\mathbf{X}_i\end{aligned}$$

所以，利用：

$$(\mathbf{Q}\mathbf{X}_i)'(\mathbf{Q}\mathbf{X}_i) = \sum_{t=1}^{T}(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)'$$

式(21.33)简化成由式(21.27)给出的估计值。

目前，许多软件都使用式(21.27)，但有一种可供选择的估计量可能会更好。特别地，序列无关误差 ε_{it} 的假设很容易被放松。若 $\boldsymbol{\varepsilon}_i$ 是 iid $[\mathbf{0},\boldsymbol{\Sigma}_i]$，则使用方差矩

阵(21.33)的更一般形式，该式满足 $\mathrm{Cov}[\mathbf{Q}\varepsilon_i,\mathbf{Q}\varepsilon_j]=\mathbf{0}$，对于 $i\neq j$，并用$(\mathbf{Q}\hat{\varepsilon}_i)(\mathbf{Q}\hat{\varepsilon}_i)'$代替 $\mathrm{V}[\mathbf{Q}\varepsilon_i]$，其中，$\hat{\varepsilon}_i=\mathbf{y}_i-\mathbf{X}_i\hat{\beta}_{\mathrm{W}}$。从而得到由式(21.28)给出的估计值。

由前面推导知，应该很明显，$\hat{\beta}_{\mathrm{W}}$ 在随机效应模型中同样是一致的，尽管正如21.7节所证明的，当随机效应模型合适，它的有效性就不如随机效应估计量那样有效。

组内估计量的 GLS 估计

组内模型(21.32)同样能通过可行 GLS 加以估计。

不过，如果事实上 ε_{it} 是 iid $[0,\sigma_\varepsilon^2]$的，那么 GLS 并没有什么益处。为了理解这一点，注意 $\mathbf{Q}\varepsilon_i$ 与 $\mathbf{Q}\varepsilon_j$ 是独立的，$i\neq j$，满足 $\mathrm{V}[\mathbf{Q}\varepsilon_i]=\sigma_\varepsilon^2\mathbf{Q}$，因此，GLS 估计量是：

$$\hat{\beta}_{\mathrm{W,GLS}}=\Big[\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{Q}'\mathbf{Q}^{-}\mathbf{Q}\mathbf{X}_i\Big]^{-1}\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{Q}'\mathbf{Q}^{-}\mathbf{Q}\mathbf{y}_i$$

其中，当 $\mathbf{Q}$ 不为满秩时，使用了广义逆 $\mathbf{Q}^{-}$。然而，对于广义逆来说，由于 $\mathbf{Q}'\mathbf{Q}^{-}\mathbf{Q}=\mathbf{Q}$，并且 $\mathbf{Q}=\mathbf{Q}\mathbf{Q}'$，当这里 $\mathbf{Q}$ 是幂等矩阵时，有 $\mathbf{Q}'\mathbf{Q}^{-}\mathbf{Q}=\mathbf{Q}'\mathbf{Q}$。在 $\hat{\beta}_{\mathrm{W,GLS}}$公式中，若用 $\mathbf{Q}'\mathbf{Q}$ 代替 $\mathbf{Q}'\mathbf{Q}^{-}\mathbf{Q}$，则得出式(21.32)中的 OLS 估计量。

如果对 ε_{it} 做出其他模型的假设，那么实施 GLS 会得到一些益处。该方法本质上与 21.5.2 节没有固定效应的混合 GLS 是一样的，只是必须剔除第一个固定效应。这就导致了非满秩的误差 $\mathbf{Q}\varepsilon_i$，因此，我们首先省略一个时期，然后将混合 OLS 应用到仅仅$(T-1)$个时期上。相反，只使用通常组内 *FE* 估计量就更容易一些，而且，往往也不缺少有效性，然后利用式(21.28)获得面板稳健标准误差。

关于短面板数据，麦柯迪(MaCurdy，1982b)曾经给出了固定效应模型 ε_{it} 的 ARMA 过程的识别与估计的 Box - Jenking 类似分析。对于短面板，不必一定要假定一个 ε_{it} 的 ARMA 过程或者甚至是平稳的，因为对于 $N\to\infty$，我们总能通过 $N^{-1}\sum_i\hat{u}_{it}\hat{u}_{is}$一致估计出 $\mathrm{Cov}[u_{it},u_{is}]$。不过，我们对决定误差的 ARMA 过程感兴趣。

21.6.2 一阶差分估计量

组内模型可借助从最初模型中减去时间平均模型 $\bar{y}_i=\alpha_i+\bar{\mathbf{x}}_i'\beta+\bar{\varepsilon}_i$ 来获得。否则，人们能减去滞后一个时期模型 $y_{i,t-1}=\alpha_i+\mathbf{x}_{i,t-1}'\beta+\varepsilon_{i,t-1}$。于是：

$$(y_{it}-y_{i,t-1})=(\mathbf{x}_{it}-\mathbf{x}_{i,t-1})'\beta+(\varepsilon_{it}-\varepsilon_{i,t-1}),\quad t=2,\cdots,T \tag{21.34}$$

因此，剔除了固定效应 α_i。进行 OLS 估计，从而得出一阶差分估计量：

$$\hat{\beta}_{\mathrm{FD}}=\Big[\sum_{i=1}^{N}\sum_{t=2}^{T}(\mathbf{x}_{it}-\mathbf{x}_{i,t-1})(\mathbf{x}_{it}-\mathbf{x}_{i,t-1})'\Big]^{-1}\sum_{i=1}^{N}\sum_{t=2}^{T}(\mathbf{x}_{it}-\mathbf{x}_{i,t-1})(y_{it}-y_{i,t-1}) \tag{21.35}$$

注意到，此回归仅有 $N(T-1)$个观测值。应用容易获得的误差是对所有 NT 个观测值加以叠放，然后减去一期滞后项。于是，仅有(1,1)观测值被省略，而所有 T 个第 1 期观测值$(i,1)$，$i=1,\cdots,N$ 必在差分滞后被剔除。

一阶差分估计量的一致性

一阶差分估计量的一致性需要 $E[\varepsilon_{it}-\varepsilon_{i,t-1}\mid \mathbf{x}_{it}-\mathbf{x}_{i,t-1}]$。这个条件比 $E[\varepsilon_{it}\mid \mathbf{x}_{it}]=0$ 强一些，却弱于组内估计量一致性所必需的强外生性条件。

一阶差分估计量的渐近分布

统计推断需要调整通常的 OLS 标准误差，以便解释误差项 $\varepsilon_{it}-\varepsilon_{i,t-1}$ 关于不同时间的相关性。为了获得 $\hat{\beta}_{\mathrm{FD}}$ 的渐近方差，将第 i 个个体的模型叠放成：

$$\Delta\mathbf{y}_i=\Delta\mathbf{X}_i'\boldsymbol{\beta}+\Delta\boldsymbol{\varepsilon}_i$$

其中，$\Delta\mathbf{y}_i$ 表示 $(T-1)\times 1$ 维向量，其元素为 $(y_{i2}-y_{i1}),\cdots,(y_{iT}-y_{i,T-1})$，$\Delta\mathbf{X}_i$ 表示 $(T-1)\times K$ 维向量，其行为 $(\mathbf{x}_{i2}-\mathbf{x}_{i1})',\cdots,(\mathbf{x}_{iT}-\mathbf{x}_{i,T-1})'$。于是，若假定对不同 i 具有独立性，则：

$$\hat{\beta}_{\mathrm{FD}}=\left[\sum_{i=1}^{N}(\Delta\mathbf{X}_i)'(\Delta\mathbf{X}_i)\right]^{-1}\sum_{i=1}^{N}(\Delta\mathbf{X}_i)'(\Delta\mathbf{y}_i) \tag{21.36}$$

具有下述方差矩阵：

$$V[\hat{\beta}_{\mathrm{FD}}]=\left[\sum_{i=1}^{N}(\Delta\mathbf{X}_i)'(\Delta\mathbf{X}_i)\right]^{-1}\left[\sum_{i=1}^{N}(\Delta\mathbf{X}_i)'V[\Delta\boldsymbol{\varepsilon}_i\mid\Delta\mathbf{X}_i](\Delta\mathbf{X}_i)\right]\left[\sum_{i=1}^{N}(\Delta\mathbf{X}_i)'(\Delta\mathbf{X}_i)\right]^{-1} \tag{21.37}$$

最简单的假设是，ε_{it} 为 iid $[0,\sigma_\varepsilon^2]$ 的。于是，误差 $(\varepsilon_{it}-\varepsilon_{i,t-1})$ 现在是 MA(1) 误差，其方差为 $2\sigma_\varepsilon^2$，而关于个体 i 的相隔一个时期自相关为 σ_ε^2。由此可得，$V[\Delta\boldsymbol{\varepsilon}_i]$ 等于 σ_ε^2 乘下述 $(T-1)\times(T-1)$ 阶矩阵：对角线上元素为 2，紧靠着对角线的非对角位置的元素为 1，而其余都为 0。

一个更现实的假设是，给定 i 时，ε_{it} 关于时间是相关的，所以对于 $t\neq s$，由 $\mathrm{Cov}[\varepsilon_{it},\varepsilon_{is}]\neq 0$，但对于不同 i 仍是独立的。由式(21.13)知，对于短面板，作为对自相关与异方差性的一般形式来说，稳健估计量是式(21.37)，该式要用 $(\widehat{\Delta\boldsymbol{\varepsilon}_i})'(\widehat{\Delta\boldsymbol{\varepsilon}_i})$ 代替 $V[\Delta\boldsymbol{\varepsilon}_i]$。人们应该永远不要使用一阶差分模型(21.37)中 OLS 回归的通常 OLS 标准误差，因为只有在 ε_{it} 为随机漫步以使 $(\varepsilon_{it}-\varepsilon_{i,t-1})$ 是 iid 的时，这样做才是正确的，但这种情况很少发生。

对于 $T=2$，由于 $\bar{y}=(y_1+y_2)/2$ 一阶差分估计量与组内估计量是相等的，因此 $(y_1-\bar{y})=(y_1-y_2)/2$，而 $(y_2-\bar{y})=-(y_1-y_2)/2$，对于 $\mathbf{x}$ 有类似情况。对于 $T>2$，这两个估计量则不一样。在最简单假设：ε_{it} 是 iid 条件下，可以证明，一阶差分模型(21.34)的 GLS 估计量等于组内估计量。估计量 $\hat{\beta}_{\mathrm{FD}}$ 反而通过对式(21.34)进行 OLS 估计，并没有 $\hat{\beta}_{\mathrm{W}}$ 那样有效。鉴于此，绝大多数引论课程不涉及一阶差分估计量。然而，若引进滞后因变量，则广泛运用一阶差分估计量（参见第 22 章）。从而，组内估计量是非一致的。一阶差分估计量也是非一致的，却依赖于允许进行一致 IV 估计的较弱外生性假设。

21.6.3 *条件 ML 估计量*

条件 MLE 是对以个体均值 $\bar{y}_1,\cdots,\bar{y}_T$ 为条件的 $y_{11},\cdots,y_{NT}$ 联合似然求极大值。此方法具有下述引人注目的特性：对于在正态性条件下的线性面板模型来说，

可剔除固定效应 α_i，所以极大化只是关于 $\boldsymbol{\beta}$ 的。

假定以回归元 $\mathbf{x}_{it}$ 为条件的 y_{it} 与参数 $\alpha,\boldsymbol{\beta},\sigma^2$ 均是 iid 的，满足正态分布 $\mathcal{N}[\alpha_i+\mathbf{x}_{it}'\boldsymbol{\beta},\sigma^2]$。于是，条件似然函数为：

$$\begin{aligned} L_{COND}(\boldsymbol{\beta},\sigma^2,\boldsymbol{\alpha}) &= \prod_{i=1}^{N} f(y_{i1},\cdots,y_{iT}\mid\bar{y}_i) \\ &= \prod_{i=1}^{N}\frac{f(y_{i1},\cdots,y_{iT},\bar{y}_i)}{f(\bar{y}_i)} \\ &= \prod_{i=1}^{N}\frac{(2\pi\sigma^2)^{-T/2}}{(2\pi\sigma^2/T)^{-1/2}} \\ &\quad\times\exp\left\{\sum_{t=1}^{T}-[(y_{it}-\mathbf{x}_{it}'\boldsymbol{\beta})^2+(\bar{y}_i-\bar{\mathbf{x}}_i'\boldsymbol{\beta})^2]/2\sigma^2\right\} \end{aligned} \tag{21.38}$$

假定对不同 i 具有独立性，第一个等式定义了条件似然。如果不用下标 i，给定 $y_1,\cdots,y_T$ 的知识，$f(y_1,\cdots,y_T\mid\bar{y})=f(y_1,\cdots,y_T,\bar{y})/f(\bar{y})$ 以及 $f(y_1,\cdots,y_T,\bar{y})=f(y_1,\cdots,y_T)$ 作为 $\bar{y}=T^{-1}\sum_i y_i$ 的信息并没有增加什么内容，所以第二个等式总是成立的。在正态性下，经过某些代数运算之后，得到第三个等式，这留作一个习题。

一个重要结果是，固定效应 α 并没有出现在式(21.38)的最后一个等式中，因此，$L_{COND}(\boldsymbol{\beta},\sigma^2,\boldsymbol{\alpha})$ 事实上就是 $L_{COND}(\boldsymbol{\beta},\sigma^2)$，我们要求条件对数函数(21.38)仅仅关于 $\boldsymbol{\beta}$ 与 σ^2 的极大值。所得到的条件 ML 估计量 $\hat{\boldsymbol{\beta}}_{CML}$ 是一阶条件：

$$\frac{1}{\sigma^2}\sum_{t=1}^{T}\sum_{i=1}^{N}[(y_{it}-\mathbf{x}_{it}'\boldsymbol{\beta})\mathbf{x}_{it}-(\bar{y}_i-\bar{\mathbf{x}}_i'\boldsymbol{\beta})\bar{\mathbf{x}}_i]=\mathbf{0}$$

的解，或等价地：

$$\sum_{t=1}^{T}\sum_{i=1}^{N}[(y_{it}-\bar{y}_i)-(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)'\boldsymbol{\beta})](\mathbf{x}_{it}-\bar{\mathbf{x}}_i)=\mathbf{0}$$

然而，这些只是出自 $(y_{it}-\bar{y}_i)$ 对 $(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)$ 的 OLS 回归的一阶条件。

因此，条件 MLE $\hat{\boldsymbol{\beta}}_{CML}$ 等于组内估计量 $\hat{\boldsymbol{\beta}}_W$。

从直观上看，此方法会得出一致估计量，因为式(21.38)中以 $\bar{y}_i$ 作为条件，剔除了固定效应。更正式地，$\bar{y}_i$ 是关于 α_i 的充分估计量，而且以充分估计量作为条件，能促使对 $\boldsymbol{\beta}$ 进行一致估计(参见 23.2.2 节)。

21.6.4 *最小二乘法虚拟变量估计量*

考察在任何差分之前的最初固定效应模型(21.22)。OLS 分析能直接用于此模型，联立估计出 $\boldsymbol{\alpha}$ 与 $\boldsymbol{\beta}$。

原则上并不需要特殊软件。人们可简单地估计出 y_{it} 对 $\mathbf{x}_{it}$ 的 OLS 回归以及一系列 N 个指示变量 $d_{1,it},\cdots,d_{N,it}$，其中，如果 $j=i$，$d_{j,it}=1$，否则为 0。然而，当 N 增大时，存在太多的回归元，使得 $(N+K)\times(N+K)$ 个回归元矩阵逆存在成为可能。不过，经过一些矩阵代数，将该问题简化成 $K\times K$ 阶矩阵的逆。可以证明，所得到的 $\boldsymbol{\beta}$ 估计量等于组内估计量。这是所谓的弗里施—沃定理(**Frisch-Waugh Throem**)关于子集合回归的特殊情况。如果虚拟变量被所有变量对虚拟变量的回

归排除，同时若源自这些回归的残差用于第二阶段回归，那么我们得到的估计值与整体回归的估计值一样。但是，此处残差是它们各自与其均值的离差，即组内回归。为了完整起见，现在阐述有关的矩阵代数运算。

一旦对所有 N 个个体的式(21.29)中的 $T\times 1$ 维向量进行叠放，就得出固定效应虚拟变量模型：

$$\begin{bmatrix}\mathbf{y}_1\\ \vdots\\ \mathbf{y}_N\end{bmatrix}=\begin{bmatrix}\mathbf{e} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \ddots & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{e}\end{bmatrix}\begin{bmatrix}\alpha_1\\ \vdots\\ \alpha_N\end{bmatrix}+\begin{bmatrix}\mathbf{X}_1\\ \vdots\\ \mathbf{X}_N\end{bmatrix}\boldsymbol{\beta}+\begin{bmatrix}\varepsilon_1\\ \vdots\\ \varepsilon_N\end{bmatrix}$$

或：

$$\mathbf{y}=[(\mathbf{I}_N\otimes\mathbf{e})\quad \mathbf{X}]\begin{bmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{bmatrix}+\varepsilon \tag{21.39}$$

其中，$\mathbf{y}$ 表示 $NT\times 1$ 维向量，克罗内克积$(\mathbf{I}_N\otimes\mathbf{e})$表示 $NT\times N$ 阶分块对角矩阵，而 $\mathbf{X}$ 表示 $NT\times K$ 阶非常值回归元矩阵。

若对这一模型进行 OLS 估计，得出最小二乘法虚拟变量估计量(**least-squares dummy variable estimator，LSDV**)：

$$\begin{bmatrix}\hat{\boldsymbol{\alpha}}_{\mathrm{LSDV}}\\ \hat{\boldsymbol{\beta}}_{\mathrm{LSDV}}\end{bmatrix}=\begin{bmatrix}(\mathbf{I}_N\otimes\mathbf{e})'(\mathbf{I}_N\otimes\mathbf{e}) & (\mathbf{I}_N\otimes\mathbf{e})'\mathbf{X}\\ \mathbf{X}'(\mathbf{I}_N\otimes\mathbf{e}) & \mathbf{X}'\mathbf{X}\end{bmatrix}^{-1}\times\begin{bmatrix}(\mathbf{I}_N\otimes\mathbf{e})'\mathbf{y}\\ \mathbf{X}'\mathbf{y}\end{bmatrix}$$

$$=\begin{bmatrix}T\mathbf{I}_N & T\bar{\mathbf{X}}\\ T\bar{\mathbf{X}}' & \mathbf{X}'\mathbf{X}\end{bmatrix}\times\begin{bmatrix}\bar{\mathbf{y}}\\ \mathbf{X}'\mathbf{y}\end{bmatrix}$$

其中，样本均值矩阵 $\bar{\mathbf{X}}=[\bar{\mathbf{x}}_1'\cdots\bar{\mathbf{x}}_N']'$，$\bar{\mathbf{x}}_i=T^{-1}\sum_{t=1}^T\mathbf{x}_{it}$，$\bar{\mathbf{y}}=[\bar{y}_1\cdots\bar{y}_N]'$，而 $\bar{y}_i=T^{-1}\sum_{t=1}^T y_{it}$。若利用分块逆公式经过某些代数运算得到：

$$\begin{bmatrix}\hat{\boldsymbol{\alpha}}_{\mathrm{LSDV}}\\ \hat{\boldsymbol{\beta}}_{\mathrm{LSDV}}\end{bmatrix}=\begin{bmatrix}\bar{\mathbf{y}}-\bar{\mathbf{X}}\,\hat{\boldsymbol{\beta}}_{\mathrm{W}}\\ [\mathbf{X}'\mathbf{X}-\bar{\mathbf{X}}'\bar{\mathbf{X}}]^{-1}(\mathbf{X}'\mathbf{y}-\bar{\mathbf{X}}'\bar{\mathbf{y}})\end{bmatrix} \tag{21.40}$$

当用求和记号对此重新表述，我们得出，由式(21.24)定义的 $\hat{\boldsymbol{\beta}}_{\mathrm{LSDV}}=\hat{\boldsymbol{\beta}}_{\mathrm{W}}$，以及由式(21.25)定义的 $\hat{\boldsymbol{\alpha}}_{\mathrm{LSDV}}=\hat{\boldsymbol{\alpha}}_{\mathrm{FE}}$，因此，LSDV 估计量等于组内估计量或固定效应估计量。

对于短面板，一个明显的潜在问题是，不能确保对 $\boldsymbol{\beta}$ 与 $\boldsymbol{\alpha}$ 的一致估计，因为存在 $N+K$ 个待估参数，并且 $N\to\infty$。值得注意的是，对 $\boldsymbol{\beta}$ 的一致估计是可行的，即使对 $\boldsymbol{\alpha}$ 是非一致估计的，除非另外 $N\to\infty$。

如果 ε_{it} 是 iid $[0,\sigma^2]$的，那么此估计量是二阶矩有效的。

由此可得，$\boldsymbol{\beta}$ 的组内估计量比其他可供选择的差分估计量更有效，差分估计量同样可剔除 α_i，诸如减去第一个观测值或前面时期的观测值。倘若误差还服从正态分布，则 LSDV 估计量等于借助通常与 OLS 等价的方法而得出 AMLE，以及具有球面正态误差的线性模型的 MLE。

21.6.5 *协方差估计量*

假定数据属于 N 个类型之一，y_{it} 在总均值 $\bar{y}$ 附近的总变异(总变化)$\sum_i\sum_t(y_{it}-\bar{y})^2$分解成组内变化$\sum_i\sum_t(y_{it}-\bar{y}_i)^2$ 与组间变化$\sum_i(\bar{y}_i-\bar{y})^2$，其中，$\bar{y}_i$ 表示

第 i 个组的均值。当组间变化增大时，隶属关系变得极为重要。为了引进回归元，协方差分析推广了该方法，在此情况下，残差平方和可类似地分解。这种框架广泛用于应用统计学中。

对于短面板，将每个个体看成一类，观测到几个时期。模型(21.3)称为协方差分析模型，因为它允许第 i 类均值随不同类而变化。这种模型的估计量即组内估计量，由此也称为协方差估计量(**covariance estimator**)。

21.7 随机效应模型

随机效应模型(21.3)能重新写成：

$$y_{it}=\mu+\mathbf{x}_{it}'\boldsymbol{\beta}+\alpha_i+\varepsilon_{it},\quad i=1,\cdots,N,\ t=1,\cdots,T \tag{21.41}$$

或者：

$$y_{it}=\mathbf{w}_{it}'\boldsymbol{\delta}+\alpha_i+\varepsilon_{it} \tag{21.42}$$

其中，$\mathbf{w}_{it}=[1\quad \mathbf{x}_{it}]$，而 $\boldsymbol{\delta}=[\mu\ \ \boldsymbol{\beta}']'$。特殊个体效应 α_i 被假定为 iid 随机变量的实现值，其分布为$[0,\sigma_\alpha^2]$，而误差是 iid $[0,\ \sigma_\varepsilon^2]$。非随机纯量截距 μ 被添加进来，与式(21.5)不同，随机效应能被正规化成具有零均值的。

否则，此模型可被看成随机系数或者变系数模型，其中唯一截距系数是随机的。这种模型能重新写成 $y_{it}=\mu+\mathbf{x}_{it}'\boldsymbol{\beta}+u_{it}$，其中，误差项 u_{it} 有两个成分 $u_{it}=\alpha_i+\varepsilon_{it}$。鉴于此，随机效应模型也称为误差成分模型，甚至更早些时候的术语可以是随机截距模型。更丰富的混合模型同样允许随机斜率，参见第 22 章。

随机效应模型存在许多一致估计量，包括：(1) 模型(21.42)的 GLS 估计；(2) 一旦假定 α_i 与 ε_{it} 均是正态分布的，模型(21.42)的 ML 估计；(3) 模型(21.42)的 OLS 估计；(4) 固定效应模型估计量，诸如组内估计量与一阶差分估计量，尽管这些估计量仅仅估计时变回归元的系数。前两个估计量是渐近等价的，但在有限样本时却依赖于 σ_α^2 与 σ_ε^2 的特定估计而变化。其余估计量是一致的，即使事实上 α_i 和 ε_{it} 为 iid 的时候它们是无效的。

21.7.1 *GLS 估计量*

μ 与 $\boldsymbol{\beta}$ 的随机效应估计量是模型(21.42)的可行 GLS 估计量，而且本节稍后将证明，它通过对变换方程：

$$y_{it}-\hat{\lambda}\bar{y}_i=(1-\hat{\lambda})\mu+(\mathbf{x}_{it}-\hat{\lambda}\,\bar{\mathbf{x}}_i)'\boldsymbol{\beta}+v_{it} \tag{21.43}$$

实施 OLS 回归而获得，其中，$v_{it}=(1-\hat{\lambda})\alpha_i+(\varepsilon_{it}-\hat{\lambda}\bar{\varepsilon}_i)$，而 $\hat{\lambda}$ 关于：

$$\lambda=1-\sigma_\varepsilon/(T\sigma_\alpha^2+\sigma_\varepsilon^2)^{1/2} \tag{21.44}$$

是一致的。等价地讲：

$$\hat{\delta}_{\mathrm{RE}}=\begin{bmatrix}\hat{\mu}_{\mathrm{RE}}\\ \hat{\boldsymbol{\beta}}_{\mathrm{RE}}\end{bmatrix}=\Big[\sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{w}_{it}-\hat{\lambda}\bar{\mathbf{w}}_i)(\mathbf{w}_{it}-\hat{\lambda}\bar{\mathbf{w}}_i)'\Big]^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{w}_{it}-\hat{\lambda}\bar{\mathbf{w}}_i)(y_{it}-\hat{\lambda}\bar{y}_i) \tag{21.45}$$

其中，$\mathbf{w}_{it}=[1\quad \mathbf{x}_{it}]$，$\mathbf{w}_i=[1\ \bar{\mathbf{x}}_i]$。一致性要求 $NT\to\infty$，通过 $N\to\infty$或 $T\to\infty$或两者全部。

若假定 ε_{it} 与 α_i 均是 iid 的，则通常源自式(21.43)的 OLS 回归的 OLS 输出能用于获得方差矩阵估计，所以：

$$\mathrm{V}\begin{bmatrix}\hat{\mu}_{\mathrm{RE}}\\ \hat{\beta}_{\mathrm{RE}}\end{bmatrix}=\sigma_\varepsilon^2\left[\sum_{i=1}^N\sum_{t=1}^T(\mathbf{w}_{it}-\hat{\lambda}\bar{\mathbf{w}}_i)(\mathbf{w}_{it}-\hat{\lambda}\bar{\mathbf{w}}_i)'\right]^{-1} \tag{21.46}$$

否则，对于短面板，利用式(21.13)可获得允许 $\alpha_i+\varepsilon_{it}$ 的具有相当一般特性的稳健方差估计。从而，得出：

$$\mathrm{V}\begin{bmatrix}\hat{\mu}_{\mathrm{RE}}\\ \hat{\beta}_{\mathrm{RE}}\end{bmatrix}=\left[\sum_{i=1}^N\sum_{t=1}^T(\tilde{\mathbf{w}}_{it}\tilde{\mathbf{w}}_{it}')\right]^{-1}\sum_{i=1}^N\sum_{t=1}^T\sum_{s=1}^T(\tilde{\mathbf{w}}_{it}\tilde{\mathbf{w}}_{is}')\hat{\tilde{\varepsilon}}_{it}\hat{\tilde{\varepsilon}}_{is}\left[\sum_{i=1}^N\sum_{t=1}^T(\tilde{\mathbf{w}}_{it}\tilde{\mathbf{w}}_{it}')\right]^{-1} \tag{21.47}$$

其中，$\tilde{\mathbf{w}}_{it}=\mathbf{w}_{it}-\hat{\lambda}\bar{\mathbf{w}}_i$，而 $\tilde{\varepsilon}_{it}=\hat{\varepsilon}_{it}-\hat{\lambda}\hat{\bar{\varepsilon}}_i$，$\hat{\varepsilon}_{it}$ 表示 RE 残差。这个估计允许 ε_{it} 的任意自相关与任何异方差性。

式(21.46)需要方差成分 σ_ε^2 与 σ_α^2。由 $(y_{it}-\bar{y}_i)$ 对 $(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)$ 的组内或固定效应回归，我们获得：

$$\hat{\sigma}_\varepsilon^2=\frac{1}{N(T-1)-K}\sum_i\sum_i((y_{it}-\bar{y}_i)-(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)'\hat{\beta}_{\mathrm{W}})^2 \tag{21.48}$$

由 $\bar{y}_i$ 对截距与 $\bar{\mathbf{x}}_i$ 的回归与具有方差为 $\sigma_\alpha^2+\sigma_\varepsilon^2/T$ 的误差方程之间的关系，我们得出：

$$\hat{\sigma}_\alpha^2=\frac{1}{N-(K+1)}\sum_i(\bar{y}_i-\hat{\mu}_{\mathrm{B}}-\bar{\mathbf{x}}_i'\hat{\beta}_{\mathrm{B}})^2-\frac{1}{T}\hat{\sigma}_\varepsilon^2 \tag{21.49}$$

得到方差成分 σ_α^2 与 σ_ε^2。σ_α^2 与 σ_ε^2 的更有效估计量是可能的[例如，参见雨宫(Amemiya，1985)]，但是这些估计量不一定提高 $\hat{\beta}_{\mathrm{RE}}$的有效性。得出更广泛的估计量是可能的。方差估计量(21.49)可能是负的，在此情况下，程序往往令 $\hat{\sigma}_\alpha^2=0$，因此，$\hat{\lambda}=0$，然后借助于混合 OLS 加以估计。

为了验证可行 GLS 估计量简化成式(21.43)的 OLS 估计，要以与固定效应模型相同的方式叠放给定 i 时所有 T 个时期的观测值。于是：

$$\mathbf{y}_i=\mathbf{W}_i\boldsymbol{\delta}+(\mathbf{e}\alpha_i+\boldsymbol{\varepsilon}_i) \tag{21.50}$$

其中，$\mathbf{y}_i$，$\mathbf{e}$，ε_i 以及 $\mathbf{X}_i$ 均在式(21.29)之后定义，而 $\mathbf{W}_i'=[\mathbf{e}\ \mathbf{X}_i']$。为了通过 GLS 进行估计，我们需要获得 $T\times1$ 维误差向量$(\mathbf{e}\alpha_i+\varepsilon_i)$的方差矩阵。给定 α_i 与 ε_{it} 的独立性，我们有 $\mathrm{E}[(\mathbf{e}\alpha_i+\varepsilon_i)(\mathbf{e}\alpha_i+\varepsilon_i)']=\mathrm{E}[\varepsilon_i\varepsilon_i']+\mathrm{E}[\alpha_i^2]\mathbf{ee}'$。由于 ε_{it} 是 iid $[0,\sigma_\varepsilon^2]$，而且 α_i 是 iid $[0,\sigma_\alpha^2]$，所以得出：

$$\boldsymbol{\Omega}=\sigma_\varepsilon^2\mathbf{I}_T+\sigma_\alpha^2\mathbf{ee}'=\sigma_\varepsilon^2\left[\mathbf{Q}+\frac{1}{\psi^2}(\mathbf{I}_T-\mathbf{Q})\right]$$

其中，$\mathbf{Q}=\mathbf{I}_T-T^{-1}\mathbf{ee}'$已在式(21.30)中引入，而 $\psi^2=\sigma_\varepsilon^2/[\sigma_\varepsilon^2+T\sigma_\alpha^2]$，一旦利用 $\mathbf{QQ}'=\mathbf{Q}$，能很容易验证 $\boldsymbol{\Omega}^{-1}=\sigma_\varepsilon^{-2}[\mathbf{Q}+\psi^2(\mathbf{I}_T-\mathbf{Q})]$，并且：

$$\boldsymbol{\Omega}^{-1/2}=\frac{1}{\sigma_\varepsilon}[\mathbf{Q}+\psi(\mathbf{I}_T-\mathbf{Q})] \tag{21.51}$$

GLS 估计量可通过任何纯量倍数的 $\mathbf{\Omega}^{-1/2}$ 左乘式(21.50)获得。现在有：

$$[\mathbf{Q}+\psi(\mathbf{I}_T-\mathbf{Q})]\mathbf{y}_i=\mathbf{y}_i-\mathbf{e}\bar{y}_i+\psi(\mathbf{y}_i-(\mathbf{y}_i-\mathbf{e}\bar{y}_i))$$
$$=\mathbf{y}_i-\lambda\mathbf{e}\bar{y}_i$$

其中，$\lambda=(1-\psi)$。一旦对式(21.50)中的 $\mathbf{W}_i$、$\mathbf{e}\alpha_i$ 以及 ε_i 实施类似代数运算，得到下述模型：

$$\mathbf{y}_i-\lambda\mathbf{e}\bar{y}_i=(\mathbf{W}_i-\lambda\mathbf{e}\overline{\mathbf{W}}_i)\boldsymbol{\delta}+(1-\lambda)\alpha_i+(\varepsilon_i-\lambda\mathbf{e}\bar{\varepsilon}_i') \tag{21.52}$$

其中，式(21.52)中的变换误差具有方差矩阵 $\sigma_\varepsilon^2\mathbf{I}_T$。GLS 估计量是式(21.52)的 OLS 估计量，但式(21.52)恰好是式(21.43)的叠放形式，只是纯量 λ 要用一致估计值来代替。

当 $T\to\infty$，斜率参数的随机效应估计量 $\hat{\beta}_{\mathrm{RE}}$ 收敛到组内估计量，从而 $\lambda\to1$。否则，经过某些代数运算，可以证明，$\hat{\beta}_{\mathrm{RE}}$ 等于组内估计量与组间估计量的矩阵加权组合。当随机效应模型合适时，此加权平均比单独利用组内估计量更能发挥作用。然而，若固定效应模型是合适的，则此加权平均是非一致的，因为组间估计量是非一致的。可以证明，截距估计量简化成 $\hat{\mu}_{\mathrm{RE}}=\bar{y}-\bar{\mathbf{X}}\hat{\beta}_{\mathrm{RE}}$。对于更详细内容，参见萧政(Hsiao，2003，第 36 页)或格林(Greene，2003)。

21.7.2 *ML 估计量*

在前一节推导中，没有假定误差的正态性。实际上，如果误差是正态的，那么我们能求对数似然关于 $\boldsymbol{\beta}$，μ，σ_ε^2 与 σ_α^2 的极大值。给定 σ_α^2 与 σ_ε^2，关于 $\boldsymbol{\beta}$ 与 μ 的 MLE 与 GLS 估计量一样，但 MLE 提供的 σ_α^2 与 σ_ε^2 不同于式(21.48)与式(21.49)给出的那些值。

因而，关于 $\boldsymbol{\beta}$ 与 μ 的 MLE 是由式(21.45)给出的，其中，$\hat{\lambda}$ 由可供选择的一致估计值 $\tilde{\lambda}=1-\tilde{\sigma}_\varepsilon/(T\tilde{\sigma}_\alpha^2+\tilde{\sigma}_\varepsilon^2)^{1/2}$ 来代替。从渐近形式上看，随机效应模型的 MLE 与 GLS 估计量是等价的，但两者在有限样本下将是不同的。

对于 MLE，或许存在两个局部极大值，而不是满足 $0<\psi^2\leqslant1$ 的似然极大值，因此为了确保全局最大值需要小心谨慎。

21.7.3 *其他估计量*

当随机效应模型是正确模型时，β 的各种不同估计量都是一致的，特别地，混合 OLS 估计量、组内估计量、一阶差分估计量以及组间估计量均是一致的。然而，如果 α_i 与 ε_{it} 都是 iid 的，那么它们是无效的，而组内估计量与一阶差分估计量只能估计时变回归元的系数。

21.8 建模问题

在本节，我们考虑线性面板数据模型中出现的某些应用问题，甚至在存在诸如内生性与滞后因变量的复杂情况下，有关专题则推迟到第 22 章。

21.8.1 混合检验

随机效应模型把所有回归参数限制成对于不同横截面与时期而言均为相同的，而固定效应模型除了截距外施加了参数不变性，截距可能随不同个体而变化。混合性检验是对这些约束的合适性进行检验。

这些检验通常是利用建立在两个线性回归中的回归元相等的检验基础上的邹检验[参见格林(Greene，2003，第130页)]，那里假定回归元具有共同的方差。依赖于对误差所做出的假设，邹检验可被应用于由OLS或GLS所估计的模型。巴尔塔基(Baltagi，2001，第4章)以及萧政(Hisao，2003，第2章)均详细分析了此问题。

对于短面板数据，不可能允许斜率参数随不同个体而变化，因为这样参数个数会趋于无穷大。然而，允许参数随时间变化。于是，将模型 $y_{it}=\gamma+\mathbf{x}_{it}'\boldsymbol{\beta}+u_{it}$ 对模型 $y_{it}=\gamma_t+\mathbf{x}_{it}'\boldsymbol{\beta}_t+u_{it}$ 进行检验。一种最明显的方法是，假定随机效应满足 $u_{it}=\varepsilon_{it}+\alpha_i$，利用随机效应GLS估计量对约束模型($\gamma_t=\gamma$ 与 $\boldsymbol{\beta}_t=\boldsymbol{\beta}$)加以估计，同时对变换模型中的约束残差平方和与无约束残差平方和进行比较。如果人们偏爱更稳健的推断，那么就应获得面板稳健标准误差，并且实施沃尔德检验。对于短面板数据，一种普通做法是，设定模型具有常值斜率参数 $\boldsymbol{\beta}$，虽然由于以时间虚拟变量作为另外的回归元，而允许截距 γ_t 随时间变化。

21.8.2 特定个体效益检验

布鲁什和帕甘(Breusch and Pagan，1980)针对存在特定个体随机效应与iid误差零假设的假设，进行对比并推导出拉格朗日乘子。这些具有仅需要出自混合OLS估计残差的辅助回归而容易实施的优点。否则，人们能假定正态性，并且进行与常值系数模型对比的随机效应MLE的似然比检验，或者进行随机效应MLE的似然比检验，或者进行随机效应模型的 $\sigma_\alpha=0$ 的沃尔德检验。

在实际应用中，人们经常拒绝常值系数模型的误差是iid的零假设。通过含有面板稳健标准误差的混合OLS，或者通过随机效应GLS加以估计。

对于短面板，存在特定个体固定效应条件下，不可能有正式检验，原因在于非主要参数问题。当只存在 NT 个观测值且 T 很小时，检验 N 个检验是否为0是不可能的。相反，21.4.3节的豪斯曼检验可用于随机效应的零假设对应于备选假设的固定效应。

21.8.3 预测

在没有个体效应的模型中，预测直接利用 $\hat{y}_{js}=\mathbf{x}_{js}'\hat{\boldsymbol{\beta}}$进行。这是对总体平均 $\mathrm{E}[y_{js}\mid\mathbf{x}_{js}]$的预测。

关于给定个体以特定个体效应为条件的情况，进行预测就更加困难。这是对 $\mathrm{E}[y_{js}\mid\mathbf{x}_{js},\alpha_i]$的预测。我们考察利用随机效应模型(21.42)关于第 i 个个体的样本外预测。于是，$y_{i,t+s}=\mathbf{w}_{it}'\boldsymbol{\delta}+u_{i,t+s}$，其中，$u_{i,t+s}=\alpha_i+\varepsilon_{i,t+s}$。一个明显预测量是用 $\hat{\boldsymbol{\delta}}_{\mathrm{RE}}$代替 $\boldsymbol{\delta}$，同时用0或 $\bar{\hat{u}}_i$ 代替 $u_{i,t+s}$，其中，$\bar{\hat{u}}_i=\bar{y}_i-\bar{\mathbf{w}}_i'\hat{\boldsymbol{\delta}}_{\mathrm{RE}}$表示关于第 i 个个体的

样本内残差平均。然而,这是无效的,因为它忽略了 $u_{i,t+s}$ 与由特定个体随机效应 α_i 诱导的样本内误差之间的相关性。此问题是 GLS 框架内而不是 OLS 框架内的更一般预测问题的例子。对于这种特殊情况,其最佳线性无偏与测量(参见 22.8.3 节)是 $\hat{y}_{i,t+s}=\mathbf{x}_{it}'\hat{\boldsymbol{\delta}}_{\mathrm{RE}}+(T\sigma_\alpha^2/(T\sigma_\alpha^2+\sigma_\varepsilon^2))\bar{\hat{u}}_i$。对于固定效应模型,一个明显预测量是 $\hat{y}_{i,t+s}=\mathbf{x}_{it}'\hat{\boldsymbol{\beta}}_{\mathrm{W}}+\hat{\alpha}_{i,\mathrm{FE}}$,但在短面板数据中,这个量再次是非一致的。

21.8.4 *双向效应模型*

到目前为止,分析都聚焦于单向模型,该模型是具有 $u_{it}=\alpha_i+\varepsilon_{it}$ 的式(21.1)。更一般模型是双向模型,满足 $u_{it}=\alpha_i+\gamma_t+\varepsilon_{it}$,并考虑到特定时间效应。于是有:

$$y_{it}=\alpha_i+\gamma_t+\mathbf{x}_{it}'\boldsymbol{\beta}+\varepsilon_{it},\quad i=1,\cdots,N,\ t=1,\cdots,T \tag{21.53}$$

此模型最初是由式(21.2)阐述的。

正如已经提及的,对于短面板数据,一种通常方法是,将特定时间效应处理成固定的,同时将它们估计成包括在回归元之中的时间虚拟的系数,依据特定个体效应是应被处理为固定的还是随机的,其分析有所不同。

如果 α_i 是固定的,γ_t 也是固定的,那么式(21.53)中 $\boldsymbol{\beta}$ 的 OLS 估计量等价于 $y_{it}-\bar{y}_i-\bar{y}_t+\bar{\bar{y}}$ 对 $\mathbf{x}_{it}-\bar{\mathbf{x}}_i-\bar{\mathbf{x}}_t+\bar{\bar{\mathbf{x}}}$ 进行回归,其中,$\bar{y}_i=T^{-1}\sum_{i=1}^{T}y_{it}$,$\bar{y}_t=N^{-1}\sum_{i=1}^{N}y_{it}$,而 $\bar{\bar{y}}=(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}y_{it}$。类似地,对 $\bar{\mathbf{x}}_i$、$\bar{\mathbf{x}}_t$ 以及 $\bar{\bar{\mathbf{x}}}$ 定义。倘若 T 很大,运用该估计方法非常方便。

相反,如果 α_i 和 γ_t 都是随机的,那么误差项将具有 γ_t 成分,而 γ_t 会引起不同个体的误差相关,然而,我们关注于对不同 i 的独立性。可以证明,GLS 估计量能通过 y_{it}^* 对常值与 $\mathbf{x}_{it}^*$ 的 OLS 回归计算出来:

$$y_{it}^*=y_{it}-\lambda_1\bar{y}_i-\lambda_2\bar{y}_t+\lambda_3\bar{\bar{y}}$$

其中,$\bar{y}_i$,$\bar{y}_t$ 以及 $\bar{\bar{y}}$ 均已经定义了,而 $\mathbf{x}_{it}^*$ 可类似于 y_{it}^* 加以定义。对于双向效应模型的这种结果与其他结果,可参见萧政(Hsiao, 2003)或巴尔塔基(Baltagi, 2001)。

21.8.5 *非平衡面板数据*

迄今为止,讨论都假定面板是平衡的,平衡意指对每个年份每一个体的数据都是可以利用的。对于不同地区的面板数据,经常是这种情况。与 i 相比,对个体的面板调查而言,经常以仍在回答调查的个体数的比例随不同时间而省略或损耗。此外,某些个体可能缺失一个或多个时期,但稍后又回来了,在一些情况下,如同轮换面板(**rotating panels**),比如 CPS 所设计的情况,一些住户被连续调查 4 个月,而有 8 个月没有调查,然后调查其他住户 4 个月。这种在不同年份出现各种不同个体的面板称为非平衡面板(**unbalanced panels**)或不完全面板(**imcomplete panels**)。

设 d_{it} 表示指示变量,当第 it 个观测值是可观测的时候,$d_{it}=1$,否则为 0。于是,对于特定个体效应模型来说,如果强外生性假设(21.4)变成:

$$\mathrm{E}[u_{it}\,|\,\alpha_i,\mathbf{x}_{i1},\cdots,\mathbf{x}_{iT},d_{i1},\cdots,d_{iT}]=0 \tag{21.54}$$

那么 FE 估计量是一致的,同时,如果 α_i 与其他条件变量是独立的,那么 RE 估计

量是一致的。从而，对固定效应估计量与随机效应估计量做相对很少的调整，就不应用于非平衡数据。这应该由作为 21.2.2 节给出的各种模型 OLS 估计量的最初估计量表示清楚地看出。例如，对于随机效应模型，用 $\hat{\lambda}_i = 1 - \sigma_\varepsilon / (T_i \sigma_\alpha^2 + \sigma_\varepsilon^2)^{1/2}$ 代替式(21.10)中的 $\hat{\lambda}$，其中，T_i 表示个体 i 的观测值个数[参见巴尔塔基(Baltagi, 1985)，万斯比克和卡普坦(Wansbeek and Kapteyn, 1989)]。戴维斯(Davis, 2002)考察了多向随机效应模型。对于固定效应模型，样本中的个体观测值必须至少有两次是观测到的，而自由度必须做出适当调整。巴尔塔基(Baltagi, 2001)对非平衡面板给出了深入细致的讨论。可以估计第 21 章至第 23 章阐述的更为标准的面板模型的经济计量软件包，通常会自动地处理缺失观测值。

有时候，通过包含样本所有年份的个体，把非平衡面板转换成平衡面板，很明显，这能大大减少有效性，原因在于损失了许多观测值。进一步地，如果数据不是随机缺失，这会恶化非代表样本的潜在问题。

缺失数据的一个原因是，尽管大多数变量是可观测的，但至少有一个变量不是可观测到的。例如，对收入问题的无回答率可以是相当高的。由于一个回归元(比如收入)的数据缺失，与其去掉全部观测值，不如利用第 27 章阐述的估算方法提高有效性。

如果从样本中去掉一些个体的原因是与误差项相关，那么非平衡面板就需要特殊方法，所以式(21.54)不会成立。例如那些具有异乎寻常低工资的个体(一旦控制可观测特性之后)可能要从面板样本中去掉。若工资是因变量，则出现非代表面板的结果，将导致损耗偏倚。一致估计要求使用推广到面板数据的样本选择方法(参见 23.5.2 节)。

21.8.6 *测量误差*

回归元的测量误差会导致横截面回归模型非一致参数估计。若使用涉及数据差分的面板数据方法，则其结果可能是增加由依赖于对数据生成过程所做假设而引起的非一致性。

21.9 应用研究

本章阐述的各种估计量都很容易实施。一种最简单的方法是，使用诸如 LLMDEP、STATA 以及 TSP 经济计量学软件包中的可用面板命令，它们均增加了具有通常处理非平衡面板的优点。否则，绝大多数估计量只要求横截面软件包对变换数据进行适当混合 OLS 回归，尽管标准误差可能不同于面板软件包标准误差，因为后者省略了由变换引起的自相关，并使用不同的自由度。

软件中面板命令的弱点是，它们目前计算的标准误差是建立在约束分布假设，诸如固定效应模型 iid 误差、随机效应模型的 iid 个体效应与 iid 误差的基础上。为了计算本章阐述的更稳健标准误差估计，需要含有面板自助法的面板估计，或利用计算聚集稳健标准误差选项的适当混合 OLS 回归。

在微观经济计量分析中，在具有固定效应的模型与没有固定效应的模型之间

存在基本差异。如果偏爱没有固定效应的模型，那么应通过豪斯曼检验来判断正确与否。若这个检验拒绝随机效应模型，那么利用下一节阐述的工具变量一致地估计时常值回归元仍是可行的。

21.10 文献注释

大部分教科书，例如格林（Greene，2003）的书，至少包括面板数据模型的章节，伍德里奇（Wooldridge，2002）中有几章内容包括线性面板模型和非线性面板模型。关于面板数据的经济计量学专题包括萧政（Hsiao，1986，2003），巴尔塔基（Baltagi，1995，2001），马加什和塞韦斯特（Matyas and Sevestre，1995），李明宰（M-J. Lee，2002）以及阿雷拉诺（Arellano，2003）。最后三本书强调了本书第22章与第23章阐述的一些方法。迪格尔、梁以及赛格尔（Diggle，Liang，and Zeger，1994，2002）则是标准的统计参考书。

21.4 芒德拉克（Mundlak，1978）撰写了固定效应与随机效应模型的经典论文。豪斯曼（Hausman，1978）运用这两个模型之间的检验去阐明他的检验方法。

21.6 库（Kuh，1959）以及奥克（Hoch，1962）提供了两个早期的面板数据在投资函数估计与生产函数估计方面的应用。这些研究都是将利用时间序列变异的组内估计值与利用横截面变异的组间变异加以比较。

习 题

21-1 [改编自巴尔塔基(Baltagi，1999)。]考察面板模型 $y_{it}=\alpha+\beta x_{it}+u_{it}$，其中，$\alpha$ 与 β 均为纯量。

(a) 证明通过适当减法，使得这个模型蕴含：

$$y_{it}-\bar{y}=\beta(x_{it}-\bar{x}_i)+\beta(\bar{x}_i-\bar{x})+(u_{it}-\bar{u})$$

其中，$\bar{y}=(NT)^{-1}\sum_{i,t}y_{it}$，$\bar{x}=(NT)^{-1}\sum_{i,t}x_{it}$，$\bar{x}_i=T^{-1}\sum_t x_{it}$。

(b) 考虑其相应的无约束最小二乘法回归：

$$y_{it}-\bar{y}=\beta_1(x_{it}-\bar{x}_i)+\beta_2(\bar{x}_i-\bar{x})+(u_{it}-\bar{u})$$

证明 β_1 的最小二乘法估计量是组内估计量，而 β_2 的最小二乘法是组间估计量。

(c) 证明当 $u_{it}=\mu_i+v_{it}$ 时，其中，$\mu_i\sim\text{iid}\,[0,\sigma_\mu^2]$，$v_{it}\sim\text{iid}\,[0,\sigma_v^2]$，而 μ_i 与 v_{it} 关于 i 和 t 都是相互独立的，OLS 与 GLS 估计量是等价的。

21-2 考察固定效应线性回归模型 $y_{it}=\alpha_i+\mathbf{x}_{it}'\boldsymbol{\beta}+\varepsilon_{it}$ 的估计，其中，α_i 是可能与 $\mathbf{x}_{it}$ 相关的固定效应。就个体 i 而言，叠放所有 T 个观测值，得到 $\mathbf{y}_i=\alpha_i\mathbf{e}+\mathbf{X}_i\boldsymbol{\beta}+\varepsilon_i$ [参见式(21.29)的定义]。考察估计量 $\hat{\boldsymbol{\beta}}=[\sum_{i=1}^N\mathbf{X}_i'\mathbf{J}'\mathbf{J}\mathbf{X}_i]^{-1}\times\sum_{i=1}^N\mathbf{X}_i'\mathbf{J}'\mathbf{J}\mathbf{y}_i$，其中，$\mathbf{J}$ 表示 $T\times T$ 阶已知常值矩阵，使得 $\mathbf{Je}=\mathbf{0}$。[注意到，$\mathbf{J}$ 的例子是 $\mathbf{Q}=\mathbf{I}_T-T^{-1}\mathbf{ee}'$。]

(a) 给出关于估计量 $\hat{\boldsymbol{\beta}}$ 的动机。

(b) 求 $\mathrm{E}[\hat{\boldsymbol{\beta}}]$。为了简单起见，假定 $\mathbf{X}_i$ 是固定回归元，而 ε_{it} 是 iid $[0,\sigma^2]$。$\hat{\boldsymbol{\beta}}$ 关于 $\boldsymbol{\beta}$ 是无偏的吗？

(c) 求 $V[\hat{\boldsymbol{\beta}}]$。为了简单起见，假定 $\mathbf{X}_i$ 是固定回归元，而 ε_{it} 是 iid $[0,\sigma^2]$。

(d) 现在假定 ε_{it} 关于 i 是独立的，但关于 t 却是与 $V[\varepsilon_i]=\Omega_i$ 相关的。求 $V[\hat{\boldsymbol{\beta}}]$。

(e) 假定效应 α_i 是随机的 $(0,\sigma_\alpha^2)$，而不是固定的。此题中估计量是一致的吗？

21-3 ［改编自巴尔塔基(Baltagi,1998)。］考察固定效应，双向误差成分面板数据模型：

$$y_{it}=\alpha+\mathbf{x}_{it}'\boldsymbol{\beta}+\mu_i+\lambda_t+\varepsilon_{it}$$

其中，α 表示纯量，$\mathbf{x}_{it}$ 表示 $K\times1$ 维内生回归元向量，$\boldsymbol{\beta}$ 表示 $K\times1$ 维向量，μ 与 λ 分别表示固定个体效应与时间效应，同时 $\varepsilon_{it}\sim$ iid $[0,\sigma^2]$。

(a) 证明 $\boldsymbol{\beta}$ 的组内估计量是最佳线性无偏的，它能通过对此模型应用两个组内(单向)变换来获得。第一个变化是忽略时间效应的组内变化，而随后忽略个体效应的组内变换。

(b) 证明这两个组内(单向)变换的次序无关紧要。给出此结果的直观解释。

21-4 利用 21.3 节工资小时数据的 50％随机子样本。

(a) 能用 β 直接解释成劳动力供给弹性吗？请解释。

(b) 对于下述估计量：(1) 混合 OLS；(2) 组间估计量；(3) 组内估计量；(4) 一阶差分估计量；(5) 随机效应 GLS；(6) 随机效应 MLE。给出(i) $\hat{\beta}$；(ii) 默认标准误差；(iii)具有 200 次复制的面板自助法标准误差。

(c) β 的估计值是相似的吗？

(d) 默认标准误差与面板稳健的标准误差之间存在系统差异吗？

(e) (b)部分固定效应模型的混合 OLS 估计量关于 β 是一致的吗？随机效应模型的混合 OLS 估计量关于 β 是一致的吗？

(f) 实施此模型中 β 的固定效应与随机效应(GLS)估计值之间差异的豪斯曼检验。这可以人工地利用前面具有默认标准误差的回归输出吗？你能得出什么结论？更偏爱哪一个模型？

(g) 给定前面证据，你认为劳动力供给曲线向上倾斜吗？请解释。

22 线性面板模型：扩展

22.1 引 论

前面几章已经阐述具有固定或随机截距而回归元为强外生的线性面板数据模型的各种变形。现在，我们转向对线性模型的各种不同推广，关注对强外生性假设的放松，以便允许对具有内生变量和/或以滞后因变量作为回归元的模型进行一致估计。

运用工具变量是处理内生回归元的标准方法。利用面板数据比利用横截面数据更容易获得工具，因为其他时期的外生回归元可用作当前时期内生回归元的工具。其唯一的复杂情况是，首先要控制任何固定或随机效应。

面板数据允许回归元额外地包括滞后因变量以及单一横截面情形的不可利用数据。这允许对下述动态模型进行估计：此动态模型可对作为不可观测特定个体效应结果(例如第 21 章所阐述的工资持久性)，与作为由先前时期结果直接决定当前时期结果而引起的持久性之间加以区分。不过，如果滞后因变量为回归元，那么控制特定个体效应的第 21 章估计量就是一致的。利用较长滞后时期项作为工具，工具变量估计就会产生一致估计。

面板数据提供了可用于估计的过剩的矩条件，这归因于拥有大量工具，并且面板模型误差通常不是 iid 的。一种自然的估计框架是面板 GMM 估计，22.2 节将对此详细阐述，而 22.3 节以对劳动力供给弹性进行估计应用进行阐明。22.4 节与 22.5 节更深入地对具有特定个体效应的以及回归元是内生的或滞后因变量的估计加以讨论。这种讨论因为可涵盖许多可能的变化形式而相当广泛。这些变形包括特定个体效应是固定的或是随机的情况、各种外生性假设，以及恰好识别的或过度识别的模型。

本章其余内容将考察其他一些独立专题，这通常不需要阅读 22.2～22.5 节内容。与面板数据模型紧密关联的一些模型，即重复横截面数据、差异中差分以及分层模型，则放在第 22.6～22.8 节阐述。

22.2 线性面板模型 GMM 估计

第 21 章的面板回归模型将纯量因变量 y_{it} 限制成只依赖同时期回归元的 $\mathbf{x}_{it}$

值，即使所有 $\mathbf{x}_{i1},\cdots,\mathbf{x}_{iT}$ 在第 21 章强外生性假设条件下潜在地成为回归元。这排除了利用来自其他时期回归元作为当前时期工具进行更有效估计的可能性。

另外，其他时期回归元可能是当前时期回归元的有效工具，该回归元或是内生变量或是因变量滞后项。因此，在强外生性假设失效导致第 21 章估计量非一致性的情况下，容易利用工具获得一致估计。

本节提供面板 GMM 估计的一般表示式，非常有用的面板 IV 估计框架自始至终地广泛用于 22.2～22.5 节。于是，我们引入非当前时期的外生变量（回归元或工具）作为工具。为了并入固定效应或随机效应，典型地包括面板模型，只要做出相对很少的改动，就能使估计建立在这种有效基础上。具体内容，推迟到下一节阐述。

22.2.1 *面板 GMM*

考察线性面板模型：

$$y_{it}=\mathbf{x}_{it}'\boldsymbol{\beta}+u_{it} \tag{22.1}$$

其中，回归元 $\mathbf{x}_{it}$ 既可能是时变成分，又可能是时常值的，可能包括截距。此处，没有特定个体效应 α_i，即放松了 22.3 节的假设，而且假定仅仅包括当前时期变量，即放松了 22.5 节的假设。假定观测值关于 i 是独立的，并假定短面板满足 T 固定且 $N\rightarrow\infty$。

以对第 i 个个体的所有 T 个观测值叠放开始，有：

$$\mathbf{y}_i=\mathbf{X}_i\boldsymbol{\beta}+\mathbf{u}_i \tag{22.2}$$

其中，$\mathbf{y}_i$ 与 $\mathbf{u}_i$ 均表示 $T\times1$ 维向量，而 $\mathbf{X}_i$ 表示 $T\times K$ 阶矩阵，其第 t 行为 $\mathbf{x}_{it}'$，因而：

$$\mathbf{y}_i=\begin{bmatrix}y_{i1}\\ \vdots\\ y_{iT}\end{bmatrix};\quad \mathbf{X}_i=\begin{bmatrix}\mathbf{x}_{i1}'\\ \vdots\\ \mathbf{x}_{iT}'\end{bmatrix};\quad \mathbf{u}_i=\begin{bmatrix}u_{i1}\\ \vdots\\ u_{iT}\end{bmatrix}$$

模型(22.2)定义了线性方程组，所以 6.9.5 节中关于具有数据独立的对不同 i 而言的系统 IV 估计的一些结果均可直接应用。

假定存在 $\mathbf{Z}_i$ 工具的 $T\times r$ 阶矩阵，其中，$r\geqslant K$ 表示工具个数，满足 r 阶矩条件：

$$\mathrm{E}[\mathbf{Z}_i'\mathbf{u}_i]=\mathbf{0} \tag{22.3}$$

建立在这些矩条件上的 GMM 估计量是求有关二次形式

$$Q_N(\boldsymbol{\beta})=\left[\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{u}_i\right]'\mathbf{W}_N\left[\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{u}_i\right]$$

的极小值，其中，$\mathbf{W}_N$ 表示 $r\times r$ 阶加权矩阵。给定 $\mathbf{u}_i=\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta}$，经过一些代数运算，得到面板 GMM 估计量(**panel GMM estimator**)：

$$\hat{\boldsymbol{\beta}}_{\mathrm{PGMM}}=\left[\left(\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{Z}_i\right)\mathbf{W}_N\left(\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{X}_i\right)\right]^{-1}\left(\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{Z}_i\right)\mathbf{W}_N\left(\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{y}_i\right)$$

此估计量一致性的根本条件是假设式(22.3)。

在许多应用中，$\mathbf{Z}_i$ 是由外生回归元的当前值与滞后值组成的。例如，假定所有回归元都是同时期外生的。于是，$\mathrm{E}[\mathbf{x}_{it}u_{it}]=\mathbf{0}$ 蕴含着 $\mathbf{Z}_i=[\mathbf{x}_{i1}' \cdots \mathbf{x}_{iT}']$。在此情况下，模型是恰好识别的，而且由于 $\mathbf{Z}_i=\mathbf{X}_i$，$\hat{\beta}_{\mathrm{PGMM}}$ 简化成第 21 章的混合估计量。如果额外假定 $\mathrm{E}[\mathbf{x}_{i,t-1}u_{it}]=\mathbf{0}$，那么 $\mathbf{x}_{it-1}$ 可用作关于第 it 个观测值的另一个工具，该模型是过度识别的(**over-identified**)，利用估计量得出更有效的估计量是可能的。

22.4 节详细阐明利用各种外生性假设来构成工具矩阵 $\mathbf{Z}_i$。当这种分析用于含有特定个体效应 α_i 的面板数据模型时，就需要加以改动。22.3 节运用一个实证应用例子对此加以阐述，而第 22.4 节与第 22.5 节则以明确方式进行讨论。

22.2.2 *面板稳健统计推断*

为了表述面板 GMM 估计量的分布，用更简洁的记号非常方便。重新写成：

$$\hat{\beta}_{\mathrm{PGMM}}=[\mathbf{X}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\mathbf{y} \tag{22.4}$$

其中，$\mathbf{X}'=[\mathbf{X}_1' \cdots \mathbf{X}_N']$，$\mathbf{Z}'=[\mathbf{Z}_1' \cdots \mathbf{Z}_N']$ 而 $\mathbf{y}'=[\mathbf{y}_1' \cdots \mathbf{y}_N']$。于是，$\hat{\beta}_{\mathrm{PGMM}}$ 是渐近正态的，其估计渐近方差矩阵为：

$$\hat{\mathrm{V}}[\hat{\beta}_{\mathrm{PGMM}}]=[\mathbf{X}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}\mathbf{W}_N(N\hat{\mathbf{S}})\mathbf{W}_N\mathbf{Z}'\mathbf{X}[\mathbf{X}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\mathbf{X}]^{-1} \text{[1]} \tag{22.5}$$

参见式(6.97)，其中，$\hat{\mathbf{S}}$ 表示 $r\times r$ 阶矩阵：

$$\mathbf{S}=\operatorname{plim}\frac{1}{N}\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{u}_i\mathbf{u}_i'\mathbf{Z}_i \tag{22.6}$$

的一致估计，同时假定关于 i 具有独立性。这里的根本假设是 $N^{-1/2}\mathbf{Z}'\mathbf{u}=N^{-1/2}\sum_i\mathbf{Z}_i'\mathbf{u}_i\xrightarrow{d}\mathcal{N}[\mathbf{0},\ S]$，$\mathbf{S}$ 的怀特形式稳健估计是：

$$\hat{\mathbf{S}}=\frac{1}{N}\sum_{i=1}^{N}\mathbf{Z}_i'\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'\mathbf{Z}_i \tag{22.7}$$

其中，$T\times 1$ 阶估计残差 $\hat{\mathbf{u}}_i=\mathbf{y}_i-\mathbf{X}_i\hat{\beta}$。

由估计式(22.5)，得出面板稳健标准(**panel-robust standard**)误差，既考察异方差性，又考虑到不同时间的相关性。否则，能使用面板自助法(**panel bootstrap**)。进一步讨论，参见 21.2.3 节对同样问题的应用。

22.2.3 *一步与两步面板 GMM*

除当 PGMM 估计量简化成关于任何 $\mathbf{W}_N$ 的 IV 估计量$[\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{Z}'\mathbf{y}$ 时恰好识别的情况之外，式(22.4)中各种不同的满秩加权矩阵 $\mathbf{W}_N$ 产生了各种不同的系统 GMM 估计量。6.4.2 节已反映出此种讨论。这里给出两个重要的 $\mathbf{W}_N$ 选择。

一步 GMM

一步 GMM 或两阶段最小二乘法估计量运用了加权矩阵 $\mathbf{W}_N=[\sum_i\mathbf{Z}_i'\mathbf{Z}_i]^{-1}=[\mathbf{Z}'\mathbf{Z}]^{-1}$，得到：

$$\hat{\beta}_{\mathrm{2SLS}}=[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} \tag{22.8}$$

---

[1] 原著该公式中的“$(N\hat{\mathbf{S}})\mathbf{W}_N'\mathbf{Z}\mathbf{X}$”应为“$(N\hat{\mathbf{S}})\mathbf{W}_N\mathbf{Z}\mathbf{X}$”，这是一个印刷错误，现已改正。——译者注

引发该估计量的动机是，可以证明，如果 $\mathbf{u}_i|\mathbf{Z}_i$ 服从 iid $[\mathbf{0},\sigma^2\mathbf{I}_T]$，那么它是建立在式(22.3)上的最优估计量。

该估计量称为一步 GMM，因为给定数据时，它可直接利用式(22.8)加以计算。不过，将它称为 2SLS，原因在于它能通过两阶段方式获得：(1) $\mathbf{X}_i$ 对 $\mathbf{Z}_i$ 的 OLS，进而得出预测 $\hat{\mathbf{X}}_i$；(2) $\mathbf{y}_i$ 对 $\hat{\mathbf{X}}_i$ 的 OLS。$\hat{\boldsymbol{\beta}}_{2SLS}$方差矩阵的估计值，关于面板和异方差性都是稳健的，这由满足 $\mathbf{W}_N=[\mathbf{Z}'\mathbf{Z}]^{-1}$的式(22.5)给出。

两步 GMM

建立在无条件矩条件(22.3)基础上的最有效 GMM 估计量运用了加权矩阵 $\mathbf{W}_N=\hat{\mathbf{S}}^{-1}$，其中，$\hat{\mathbf{S}}$ 表示关于 $\mathbf{S}$ 是一致的，这已由式(22.6)定义了；一般结果，参见 6.4.2 节。一旦使用式(22.7)中的 $\hat{\mathbf{S}}$，则得出两步 GMM 估计量：

$$\hat{\boldsymbol{\beta}}_{2SGMM}=[\mathbf{X}'\mathbf{Z}\hat{\mathbf{S}}^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}\hat{\mathbf{S}}^{-1}\mathbf{Z}'\mathbf{y} \tag{22.9}$$

于是，式(22.5)得以简化，并且 $\hat{\mathrm{V}}[\hat{\boldsymbol{\beta}}_{2SGMM}]=[\mathbf{X}'\mathbf{Z}(N\hat{\mathbf{S}})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}$。

这个估计量称为两步 GMM，因为 $\boldsymbol{\beta}$ 的第一步一致估计估计量比如$\hat{\boldsymbol{\beta}}_{2SLS}$，需要用于计算残差 $\hat{\mathbf{u}}_i$，而 $\hat{\mathbf{u}}_i$ 则用于计算 $\hat{\mathbf{S}}$。

提高有效性

在本章，关注于 $\mathbf{Z}$ 不能包括 $\mathbf{X}$ 的所有成分，因为 $\mathbf{X}$ 的一些成分具有内生性。为了理解这一点，假定 $\mathbf{X}$ 是强外生的。若令 $\mathbf{Z}=\mathbf{X}$，两步 GMM 估计量简化成 $[\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{y}$，从而对面板 GMM 而言没有什么益处。然而，如果 $\mathbf{Z}$ 等于 $\mathbf{X}$ 以及另一些变量，诸如回归元的幂或者不同于当前时期的其他时期回归元值，那么两步 GMM 方法至少与 OLS 一样有效，若误差 u_{it} 服从 iid 的，则等式成立。

获得比$\hat{\boldsymbol{\beta}}_{2SGMM}$更为有效的估计量是可能的，这要借助于放松 $\mathbf{Z}_i$ 的定义，通过利用基于 $\mathrm{E}[\mathbf{u}_i|\mathbf{Z}_i]=\mathbf{0}$ 的最优矩条件，它不必是 $\mathrm{E}[\mathbf{Z}_i'\mathbf{u}_i]=\mathbf{0}$(参见 22.4.3 节)，同时利用另外矩约束。我们避开了将两步 GMM 称为最优 GMM 估计量，正如 6.3 节一样，它仅在给定式(22.3)时为最优的。

检验过度识别约束

如果存在 r 个工具，且仅有 K 个待估参数，那么面板 GMM 估计留下$(r-K)$个过度约束。由 6.3.8 节知，这使得检验过度识别约束

$$\mathrm{OIR}=\left[\sum_{i=1}^{N}\hat{\mathbf{u}}_i'\mathbf{Z}_i\right](N\hat{\mathbf{S}})^{-1}\left[\sum_{i=1}^{N}\mathbf{Z}_i'\hat{\mathbf{u}}_i\right] \tag{22.10}$$

成为可能，其中，$\hat{\mathbf{u}}_i=\mathbf{y}_i-\mathbf{Z}_i'\hat{\boldsymbol{\beta}}_{2SGMM}$，$\hat{\mathbf{S}}$ 已由式(22.7)给出，同时假定对不同 i 具有独立性，却允许给定 i 时关于不同 t 具有异方差性及相关性。注意到，必须使用$\hat{\boldsymbol{\beta}}_{2SGMM}$，而不是 $\hat{\boldsymbol{\beta}}_{2SLS}$。

在零假设：过度识别约束是有效的条件下，这一检验统计量服从 $\chi^2(r-K)$分布。当 QIR 很大时，就要拒绝过度矩条件，从而我们得出结论：$\mathbf{Z}_i$ 的一些工具与误差是相关的，从而这些工具是内生的。

22.2.4 *选取工具*

迄今为止，讨论都假定存在满足式(22.3)的 $T\times r$ 阶工具 $\mathbf{Z}_i$ 矩阵。现在，我们给出深入细致的讨论，阐明如何在面板背景下获得工具。

在横截面模型中，内生变量可借助于关注方程中没有出现的作为回归元的工具。这类变量同样能用作面板情况下的工具。然而，就面板模型而言，其他时期的数据提供了额外矩条件以及额外工具，这很容易导致 $\boldsymbol{\beta}$ 的识别或过度识别。

当对 u_{it} 与 $\mathbf{z}_{it}$ 之间相关性做出逐渐增强的一些假设时，矩条件以及利用工具个数可得以推广，其中 $s,t=1,\cdots,T$。我们遵循李明宰(M.-J. Lee, 2002)的线索，考察逐渐增强外生性假设的效果，可参见 2.3 节。强调内容不止一次地利用回归元的外生成分作为工具，但该方法还可应用于成为排除于回归(22.1)之外的变量的更传统工具。

求和假设

一种明显的方法是，类似于 $\mathbf{X}_i$ 去定义 $\mathbf{Z}_i$。于是：

$$\mathbf{Z}_i=\begin{bmatrix}\mathbf{z}_{i1}'\\ \mathbf{z}_{i2}'\\ \vdots\\ \mathbf{z}_{iT}'\end{bmatrix},\quad \mathbf{u}_i=\begin{bmatrix}u_{i1}\\ u_{i2}\\ \vdots\\ u_{iT}\end{bmatrix} \tag{22.11}$$

其中，$\mathbf{z}_{it}$ 是 $r\times 1$ 维的，如果求和假设：

$$\mathrm{E}\left[\sum_{t=1}^{T}\mathbf{z}_{it}u_{it}\right]=\mathbf{0} \tag{22.12}$$

得到满足，那么 $\mathrm{E}[\mathbf{Z}_i'\mathbf{u}_i]=\mathbf{0}$。

如果式(22.12)中 $\mathbf{z}_{it}=\mathbf{x}_{it}$，那么由式(22.4)定义的 PGMM 估计量简化成 $(\sum_i\mathbf{Z}_i'\mathbf{X}_i)^{-1}\sum_i\mathbf{Z}_i'\mathbf{y}_i$，所以这个求和假设可用于 y_{it} 对 $\mathbf{x}_{it}$ 的混合 OLS 回归。

为此估计量成为可行的，至少需要满足阶条件，因此 $r\geqslant K$。在求和假设下，寻找面板数据的工具就如同横截面数据一样困难。

同时期外生假设

一个较强的且更自然的假设是同时期外生性假设(**contemporaneous exogeneity assumption**)，即：

$$\mathrm{E}[\mathbf{z}_{it}u_{it}]=\mathbf{0},\quad t=1,\cdots,T \tag{22.13}$$

因此，假定工具与误差项同时期不相关。

这种表述引出更多矩条件，原则上与 Tr 个矩条件一样多，其中，$r=\dim[\mathbf{z}_{it}]$。为了运用这些矩条件，我们定义：

$$\mathbf{Z}_i=\begin{bmatrix}\mathbf{z}_{i1}' & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \mathbf{z}_{i2}' & & \vdots\\ \vdots & & \ddots & \mathbf{0}\\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{z}_{iT}'\end{bmatrix},\quad \mathbf{u}_i=\begin{bmatrix}u_{i1}\\ u_{i2}\\ \vdots\\ u_{iT}\end{bmatrix} \tag{22.14}$$

其中，$\mathbf{Z}_i$ 现在表示 $T\times Tr$[1]。矩条件(22.3)成立，因为由式(22.13)知 $\mathrm{E}[\mathbf{Z}_i'\mathbf{u}_i]=\mathbf{0}$，

〔1〕 原文这里为“$Tr\times T$”，应为“$T\times Tr$”，这可能是一个印刷错误，现已改正。——译者注

但现在式(22.3)定义了可用于估计 β 的 K 个分量的 Tr 矩条件。

由于隐性假设:β 是时常值的,所以才会出现矩约束明显过度的显著结果,因此,每一个额外时期均贡献额外矩约束。

额外矩约束的个数简化成 β 为时变的程度。特别地,借助于 $\mathbf{x}_{it}$ 包括$(T-1)$个时间虚拟变量,当 $t=s$ 时,则 $d_{s,it}=1$,否则 $d_{s,it}=0$,对于 $s=2,\cdots,T$,往往允许截距随时间变化。于是,不能使用条件 $\mathrm{E}[d_{s,it}u_{it}]=0$,因为它重复了包含 $\mathbf{x}_{it}$ 中的一个截距的条件 $\mathrm{E}[1\times u_{it}]=0$。在前面例子中,如果 $\mathbf{x}_{it}$ 包含时间虚拟变量,那么就只存在 $TK-(T-1)$个可利用的矩条件。任何时常值回归元只能用作一次工具。

弱外生性假设

矩条件(22.13)仅仅考察工具与回归元之间的同时期相关。一个较强的假设是弱外生性假设(**weak exogeneity assumption**)或先决工具假设(**predertermined instruments assumption**),该假设还包括工具的滞后值与当前误差是不相关的,所以:

$$\mathrm{E}[\mathbf{z}_{is}u_{it}]=\mathbf{0}, \quad s\leqslant t,\ t=1,\cdots,T \tag{22.15}$$

条件(22.15)允许 $\mathbf{z}_{i1},\cdots,\mathbf{z}_{it}$ 成为 u_{it} 的工具,尽管不能使用 $\mathbf{z}_{is}$ 的未来值。工具 $\mathbf{Z}_i$ 在构造上类似于式(22.14),只是 $\mathbf{z}_{it}'$ 要用扩展工具向量$[\mathbf{z}_{i1}',\cdots,\mathbf{z}_{it}']$来代替,该工具向量会随着 t 增大而增大。

理性预期模型以及在不确定性条件下的跨期决策模型,都会产生欧拉条件 $\mathrm{E}[u_{it}\mid\mathcal{I}_{it}]=0$,其中,$\mathcal{I}_{it}$ 表示在时间 t 时可利用的信息集合,而 u_{it} 的例子已由 6.2.7 节给出。如果信息集合包括 $\mathbf{z}_{it}$ 当前值及过去值,那么 $\mathrm{E}[u_{it}\mid\mathbf{z}_{is}]=0$, $s\leqslant t$,从而得到式(22.15)。

更一般地,这些条件在含有滞后因变量作为回归元的动态模型里是有意义的(参见 22.5 节)。在一些例子中,同时期相关并没有被排除,因而式(22.15)中的不等式 $s\leqslant t$ 要用 $s<t$ 代替。

注意到,时常值工具只能使用一次。因而,当 $\mathbf{z}_{it}=[\mathbf{z}_{1i}\ \mathbf{z}_{2it}]$时,$\mathbf{z}_{1i}$ 与 $\mathbf{z}_{2i1},\cdots,\mathbf{z}_{2it}$ 都可作为工具。

强外生性假设

一个比弱外生性更强的假设是强外生性假设(**strong exogeneity assumption**),即指工具的未来值也与当前时期误差是不相关的,因此:

$$\mathrm{E}[\mathbf{z}_{is}u_{it}]=\mathbf{0}, \quad s,t=1,\cdots,T \tag{22.16}$$

于是,$\mathbf{z}_{is}$ 的当前值、过去值以及未来值均是 u_{it} 的有效工具。

该假设对于第 21 章全部的回归元来说都要成立,因为 $\mathrm{E}[u_{it}\mid\mathbf{x}_{i1},\cdots,\mathbf{x}_{iT}]=0$ 蕴含着 $\mathrm{E}[u_{it}\mid\mathbf{x}_{is}]=0, 1\leqslant s\leqslant T$,从而 $\mathrm{E}[\mathbf{x}_{is}\mathbf{u}_{it}]=\mathbf{0}$。就静态模型而言,它是合适的,但对于动态模型来说,至多假定工具的弱外生性。

条件(22.16)允许 $\mathbf{z}_{i1},\cdots,\mathbf{z}_{iT}$ 成为 u_{it} 的工具。工具 $\mathbf{Z}_i$ 在构造上类似于式(22.14),只是式(22.14)中的 $\mathbf{z}_{it}'$ 用扩展工具向量$[\mathbf{z}_{i1}',\cdots,\mathbf{z}_{iT}']$代替。

就弱外生性情况而论,时常值工具只能利用一次。若 $\mathbf{z}_{it}=[\mathbf{z}_{1i}\ \mathbf{z}_{2it}]$,则可以利用 $T(r_{\mathrm{TI}}+r_{\mathrm{TV}})$矩条件,其中,$r_{\mathrm{TI}}$与 r_{TV}表示时常值与时变工具的数目。

矩条件数目极多,多到与 rT^2 一样,原因在于面板模型(22.1)隐含地做出排除

性约束。为了简单起见，假定 $\mathbf{x}_{it}$ 的所有成分都是强外生的，并且如有可能，我们希望使用这些作为工具。通常，y_{it} 在所有时期都依赖于回归元 $\mathbf{x}_{i1},\cdots,\mathbf{x}_{iT}$。与之相比，满足 $\mathrm{E}[\mathbf{x}_{it}u_{it}]=\mathbf{0}$ 的面板模型 $y_{it}=\mathbf{x}_{it}'\boldsymbol{\beta}+u_{it}$ 在此 y_{it} 的模型中只包含了 $\mathbf{x}_{it}$。于是，除 $\mathbf{x}_{it}$ 之外，强外生性假设即 $\mathrm{E}[\mathbf{x}_{is}u_{it}]=\mathbf{0}$ 允许排除回归元 $\mathbf{x}_{is}$，$s\neq t$ 用于工具。

冗余工具

如果 $\mathbf{z}_{it}$ 既随 i 变化又随 t 变化，那么 $\mathbf{z}_{it}$ 的滞后项与前置项也可用作工具，但这要依赖于做出的外生性假设。对于第 it 个观测值来说，可利用工具在同时期外生性下是 $\mathbf{z}_{it}$，在弱外生性下是 $\mathbf{z}_{i1},\cdots,\mathbf{z}_{it}$，而在强外生性下则是 $\mathbf{z}_{i1},\cdots,\mathbf{z}_{iT}$。这使得只利用外生回归元作为工具进行识别成为可能。与那些横截面情况相比，寻求有效工具的困难在于只有求和假设。

不过，在实际应用中，并不存在前文所述那样多的可利用工具。时常值工具(**time-invariant instruments**)$\mathbf{z}_{it}=\mathbf{z}_i$ 只能利用一次，从而对于所有 s 与 t，有 $\mathbf{z}_{it}=\mathbf{z}_{is}$。例如，这就是截距或种族或性别指示变量的情况。若工具是模型中出现的回归元与回归元滞后值，则利用工具数目就会减少。在所有时期，也许不能利用以某种系统方式变化的时变工具。因而，如果使用时间虚拟变量的完整集合，就应该包括作为工具的时间虚拟变量与时常值回归元之积。一些例子包括，时间虚拟变量、时间虚拟变量与种族或性别交互作用的指示变量。作为时间线性函数的工具应该只能利用一次。例如，如果年份是工具，就不应该再使用滞后年份。这种评论的确不可用于年龄，这对每个个体而言会以线性方式增大，却随不同个体而变化。

很明显，使用冗余工具很容易疏忽细节。若仍存在充足的非冗余工具，则面板 GMM 估计量仍然是可行的，同时通常结果是有效的。例如，如果有 r 个工具可以使用，并且其中有两个为冗余的，倘若 $r\geqslant K+2$，当 $\mathbf{Z}'\mathbf{X}$ 还是满秩的且为 K，那么该模型就是可估计的。如果使用太多的冗余工具，那么可能产生 GMM 估计的奇异性问题。即使模型是过度识别的，当一些工具是冗余的时候，过度识别约束检验的自由度将会减少。

弱工具

弱工具已在 4.9 节引入，但是不要与弱外生性相混淆。弱工具的正式检验还没有很好地建立起来。标准统计量诊断已经由 4.9 节给出。增加工具解释力至关重要。因此，控制外生回归元的偏 R^2 同样处于应该使用的工具集合中。此外，鉴于内生回归元对所有工具进行回归，统计量应是没有成为外生回归元工具子集的整体显著的代表。

由于这里的误差不是 iid 的，所以 F 统计量应该建立在面板稳健标准误差基础上。它被计算成 W/r^*，其中，W 表示由 7.2.7 节给出的排除性约束的沃尔德卡方检验统计量，而 r^* 表示那种不是最初模型中回归元的工具数目。

22.2.5 *面板 GMM 估计量的计算*

上面一节讨论的矩条件提供了工具矩阵 $\mathbf{Z}_i$。于是，给定 $\mathbf{Z}_i$，人们能通过式(22.8)定义的 $\hat{\boldsymbol{\beta}}_{2SLS}$ 或式(22.9)定义的 $\hat{\boldsymbol{\beta}}_{2SGMM}$ 估计 $\boldsymbol{\beta}$。

与两步 GMM 相比，更容易实施 2SLS 估计量。考察求和假设下的估计，其中，

$\mathbf{Z}_i$ 已由式(22.11)定义。于是，$\hat{\beta}_{2SLS}$ 由式(22.8)给出，其中，$\mathbf{Z}'\mathbf{X}=\sum_i \mathbf{Z}_i'\mathbf{X}_i=\sum_i \sum_t \mathbf{z}_{it}\mathbf{x}_{it}'$，同时类似代数运算可用于其他叉积。这就得出标准教科书中的 2SLS 公式，只是求和既关于 i 又关于 t 而进行。因而，一旦利用横截面软件包，$\hat{\beta}_{2SLS}$ 可通过 y_{it} 对 $\mathbf{x}_{it}$ 的回归来获得。于是，面板稳健标准误差能利用下述方式来获得；即利用允许对 i 聚集的聚集稳健选项，或者通过对 i 而不是既对 i 又对 t 重复抽样的面板自助法。这些方法类似于由 21.2.3 节给出的混合 LS，那里提供了额外详情。

对于不是求和假设的假设来说，人们仍然能通过适当定义工具矩阵 $\mathbf{Z}_i$，使用横截面 2SLS 软件包，从而拥有更为复杂的形式。就同时期外生性假设而言，$\mathbf{Z}_i$ 是由式(22.14)定义的。如果式(22.11)中的第 t 行 $\mathbf{z}_{it}'$ 由

$$[\mathbf{0}_{r_1}' \cdots \mathbf{0}_{r_{t-1}}' \ \mathbf{z}_{it}' \ \mathbf{0}_{r_{t+1}}' \cdots \mathbf{0}_{r_T}'] \tag{22.17}$$

代替，那么这与式(22.11)的形式相同，其中，$r_s=\dim[\mathbf{z}_{is}]$ 而 $\mathbf{0}_{r_s}$ 表示零向量。类似地，对于弱外生性假设来说，$\mathbf{z}_i$ 由式(22.11)定义，式(22.11)中的第 t 行 $\mathbf{z}_{it}'$ 由

$$[\mathbf{0}_{r_1}' \cdots \mathbf{0}_{r_{t-1}}' \ (\mathbf{z}_{it}^t)'\mathbf{0}_{r_{t+1}}' \cdots \mathbf{0}_{r_T}'] \tag{22.18}$$

代替；其中，$(\mathbf{z}_{it}^t)'=[\mathbf{z}_{i1}' \cdots \mathbf{z}_{it}']$，而 $r_s=\dim[\mathbf{z}_{is}^s]$，另外对于强外生性假设来说，$\mathbf{Z}_i$ 由式(22.11)定义，式(22.11)中的第 t 行 $\mathbf{z}_{it}'$ 由

$$[\mathbf{0}_{r_1}' \cdots \mathbf{0}_{r_{t-1}}' \ (\mathbf{z}_{it}^T)'\mathbf{0}_{r_{t+1}}' \cdots \mathbf{0}_{r_T}'] \tag{22.19}$$

代替，其中，$(\mathbf{z}_{it}^T)'=[\mathbf{z}_{i1}' \cdots \mathbf{z}_{iT}']$，而 $r_s=\dim[\mathbf{z}_{is}^T]$。生成工具的实际例子将由 22.3 节给出。

实际应用中，存在太多的矩条件。例如，含有 10 个时期数据与 5 个时变回归元，其强外生性假设会产生 500(5×10^2)个矩条件(而且前面的行向量拥有 500 个元素)，仅有 5 个要估计的参数。工具的临界值可以是非常轻微的，因为工具之间不断增加的多重共线性导致了弱工具的情形。好的实践做法是把随时间稍微变化的时变工具处理成时常值的。例如，仅仅利用第一个时期作为工具。甚至随时间变化相当大的工具可能仅仅使用几个时期而不是所有可能时期。

只利用软件包获得更有效计算，这是不可能的。相反，要么需要更专门化的软件，要么需要利用矩阵语言算法对估计量加以编程。

表 22.1 提供了四种外生性假设并概括了所得到的有效工具。

表 22.1 面板外生性假设与得到的工具

外生性假设	矩条件	工具向量[a]
求和假设	$\mathrm{E}[\sum_t \mathbf{z}_{it}u_{it}]=\mathbf{0}$	$[\mathbf{z}_{it}]$
同时期假设	$\mathrm{E}[\mathbf{z}_{it}u_{it}]=\mathbf{0}$，所有 t	$[\mathbf{0}_{r_1}' \cdots \mathbf{0}_{r_{t-1}}' \ \mathbf{z}_{it}'\mathbf{0}_{r_{t+1}}' \cdots \mathbf{0}_{r_T}']$
弱假设	$\mathrm{E}[\mathbf{z}_{is}u_{it}]=\mathbf{0}$，$s\leqslant t$，所有 t	$[\mathbf{0}_{r_1}' \cdots \mathbf{0}_{r_{t-1}}' \ (\mathbf{z}_{it}^t)'\mathbf{0}_{r_{t+1}}' \cdots \mathbf{0}_{r_T}']$
强假设	$\mathrm{E}[\mathbf{z}_{is}u_{it}]=\mathbf{0}$，所有 s 与 t	$[\mathbf{0}_{r_1}' \cdots \mathbf{0}_{r_{t-1}}' \ (\mathbf{z}_{it}^T)'\mathbf{0}_{r_{t+1}}' \cdots \mathbf{0}_{r_T}']$

[a] 工具向量是式(22.11)中 $\mathbf{Z}_i$ 的第 i 行；$(\mathbf{z}_{it}^t)'=[\mathbf{z}_{i1}' \cdots \mathbf{z}_{iT}']$，$(\mathbf{z}_{it}^T)'=[\mathbf{z}_{i1}' \cdots \mathbf{z}_{iT}']$；而 $r_s=\dim[\mathbf{z}_{is}']$ 或 $\dim[\mathbf{z}_{is}^t]$ 或 $\dim[\mathbf{z}_{is}^T]$。

22.2.6 估计的变分

尽管 $\hat{\boldsymbol{\theta}}_{2SGMM}$ 比 $\hat{\boldsymbol{\theta}}_{2SLS}$ 更为有效，一些研究发现，它具有比 $\hat{\boldsymbol{\theta}}_{2SLS}$ 更大的有限样本偏倚，尤其是当 r 非常大于 K 时。为了解释，请参见 6.3.5 节对最优 GMM 有限样本偏倚的讨论。

一种明智方法是使用工具，尽管因为加入额外工具而损失了潜在的有效性。

几位作者已经提出可供选择的 GMM 估计量，该估计量在有限样本中可能较少是有偏的。6.4.4 节已经对这样一些估计量加以讨论过，而齐利亚克（Ziliak，1997）在面板研究中使用了这样的估计量。

22.2.7 张伯伦最优距离估计量

考察特定个体效应模型的估计：

$$y_{it}=\alpha_i+\mathbf{x}_{it}'\boldsymbol{\beta}+u_{it} \tag{22.20}$$

此时回归元是强外生的，如同第 21 章一样。21.2.3 节与 21.6.1 节已经讨论了，获得组内估计量面板稳健标准误差的方法。

如果实施面板稳健推断是必要的，由于 ε_{it} 不是 iid 的，那么第 21 章所述估计量实际上都是无效的。更有效的估计可能是将最优 GMM 用于过度识别模型。当额外工具与 GMM 能应用于变换模型时，如果消除 α_i 是必要的，这里可以利用 $\mathbf{x}_{is}$，$s\neq t$（参见 22.4.2 节）。其有效性改进类似于含有异方差性的横截面数据（参见 6.3.5 节）。

张伯伦（Chamberlain，1982，1984）曾提出下述更有效的估计量。对模型（22.20）进行叠放表示，得到：

$$\mathbf{y}_i=\mathbf{e}\alpha_i+(\mathbf{I}_T\otimes\boldsymbol{\beta}')\mathbf{x}_i+\mathbf{u}_i \tag{22.21}$$

其中，$\mathbf{e}=(1,1,\cdots,1)'$ 表示 $T\times1$ 维单位向量，$\mathbf{x}_i=[\mathbf{x}_{i1}',\cdots,\mathbf{x}_{iT}']$ 表示 $TK\times1$ 维向量，而 $\mathbf{y}_i$ 与 $\mathbf{u}_i$ 表示 $T\times1$ 维向量。式(22.21)使得下面情况清楚可见：即设定 y_{it} 只依赖于同时期 $\mathbf{x}_{it}$ 的静态模型隐含做出一些约束。张伯伦使用了依赖于比条件期望的那些假设更弱的假设的线性投影推理。设：

$$\mathrm{E}^*[\alpha_i|\mathbf{x}_i]=\mu+\sum\nolimits_t\boldsymbol{\lambda}_t'\mathbf{x}_{it}=\mu+\boldsymbol{\lambda}'\mathbf{x}_i$$

其中，E^* 表示线性投影。一旦给定 $\mathrm{E}[\mathbf{u}_i|\alpha_i,\mathbf{x}_i]=\mathbf{0}$，式(22.21)蕴含着：

$$\mathrm{E}^*[\mathbf{y}_i|\mathbf{x}_i]=\mathbf{e}\mu+(\mathbf{I}_T\otimes\boldsymbol{\beta}'+\mathbf{e}\boldsymbol{\lambda}')\mathbf{x}_i$$

这对无约束线性 $\mathrm{E}^*[\mathbf{y}_i|\mathbf{x}_i]=\boldsymbol{\pi}_0+\boldsymbol{\pi}'\mathbf{x}_i$ 投影施加了约束，具体来说，$\boldsymbol{\pi}-\mathbf{I}_T\otimes\boldsymbol{\beta}'+\mathbf{e}\boldsymbol{\lambda}'=\mathbf{0}$。

张伯伦没有使用 GMM，而是提出下面两步方法。首先，通过 $\mathbf{y}_i$ 对截距与 $\mathbf{x}_i$ 进行多变量 OLS 回归获得。其次，获得求

$$Q_N(\boldsymbol{\beta},\boldsymbol{\lambda})=(\mathrm{Vec}[\hat{\boldsymbol{\pi}}-\mathbf{I_T}\otimes\boldsymbol{\beta}'-\mathbf{e}\boldsymbol{\lambda}'])'\mathbf{W}_N(\mathrm{Vec}[\hat{\boldsymbol{\pi}}-\mathbf{I_T}\otimes\boldsymbol{\beta}'-\mathbf{e}\boldsymbol{\lambda}'])$$

极小值的最优 MD 估计量（参见 6.7 节），其中，最优加权矩阵 $\mathbf{W}_N=(\hat{\mathrm{V}}[\mathrm{Vec}[\hat{\boldsymbol{\pi}}]])^{-1}$。如果 u_{it} 是异方差的，这就产生了比式(22.20)的 OLS 估计更有效的估计量 $\hat{\boldsymbol{\beta}}$。

最小距离估计被 GMM 取代；参见阿雷拉诺（Arellano，2003，第 22～23 页），以及克雷蓬和梅尔斯（Crépon and Mairesse，1995）对张伯伦的 MD 估计量与 GMM 的比较。然而，张伯伦的通过外生性假设与关于个体效应假设而获得的矩约束方法对面板文献产生了巨大的影响。他的 MD 估计量同样可用于协方差结构的估计（参见 22.5.4 节）。

22.3 面板 GMM 例子：小时与工资

我们回到 21.3 节的小时工资例子上。与第 21 章不同，现在允许回归元是内生的，并且与 22.2 节不一样，包括特定个体固定效应。在一阶差分剔除固定效应之后，通过 22.2 节方法加以估计。

回归模型是：

$$\mathrm{lnhrs}_{it}=\alpha_i+\beta_1\mathrm{lnwg}_{it}+\beta_2\mathrm{kids}_{it}+\beta_3\mathrm{age}_{it}+\beta_4\mathrm{agesq}_{it}+\beta_5\mathrm{disab}_{it}+u_{it}$$

其中，关注内容在于劳动力供给的跨期替代工资弹性 β_1，即 lnwg 的系数，并且回归元分别是孩子的数量、年龄、年龄平方以及无能力的指示变量。

麦柯迪（McCurdy，1981）在不确定条件下，利用生命周期供给模型推导出这种关系。于是，此模型就是“λ 常值”模型，其中，这里的 α_i 等于 λ_i，即最初财富的边际效用倍数是时常值的，但随不同个体而变化。由于 λ_i 依赖于变量与约束，从而需要将它处理成固定效应而不是随机效应。

22.4.2 节将进一步讨论的一种方法是，对回归方程进行一阶差分，得出：

$$\Delta\mathrm{lnhrs}_{it}=\beta_1\Delta\mathrm{lnwg}_{it}+\beta_2\Delta\mathrm{kids}_{it}+\beta_3\Delta\mathrm{age}_{it}+\beta_4\Delta\mathrm{agesq}_{it}+\beta_5\Delta\mathrm{disab}_{it}+\Delta u_{it} \tag{22.22}$$

如果所有回归元是外生的，那么利用 OLS 得到的估计值关于 $\boldsymbol{\beta}$ 是一致的。注意到，虽然 u_{it} 是 iid 的，但这种差分引起了误差序列相关，因此，应使用面板稳健标准误差。

不过，齐利亚克（Ziliak，1997）允许 lnwg_{it} 与 u_{it} 成为同时期相关的，原因在于工资测量误差或预算约束有结点。从而，式(22.22)的 OLS 估计量是非一致的。

齐利亚克提出利用合适滞后回归元作为工具的 IV 估计。假定过去工资与误差是无关的，因此，除与误差是同时期相关之外，lnwg 是弱外生的。于是，对于 $s\leqslant t-1$，$\mathrm{E}[\mathrm{lnwg}_{is}u_{it}]=0$，蕴含着差分模型误差 $\mathrm{E}[\mathrm{lnwg}_{is}\Delta u_{it}]=0$，对于 $s\leqslant t-2$，所以滞后两时期或多时期可用作一阶差分模型的工具。注意到，这意味着为了识别 $\boldsymbol{\beta}$，至少需要三个时期的最初数据。

齐利亚克的研究关注于含有内生回归元的面板 GMM 估计量的性质，因此，他将式(22.22)的所有回归元处理成内生的，并用作其他四个回归元中滞后一个或多个时期的工具。为了简单起见，截距与时间虚拟变量，以及只使用一次的时常值个

体工具都没有包括在内。就包含截距而言,此结果变化很小,因为因变量是差分形式。由于 $\ln wg_{i,t-2}$ 总是用作工具,所以前两年数据被省略,而仅有 1981～1988 年的 8 年期间数据用于估计式(22.22)。

表 22.2 阐述了由齐利亚克(Ziliak, 1997)的表 1 与表 2 给出许多结果的一个子集。为了完整起见,已经给出各种标准误差估计值,但应使用面板稳健标准误差。

表 22.2　小时与工资:线性面板模型估计量[a]

		基准情况		叠放情况	
	OLS	2SLS	2SGMM	2SLS	2SGMM
β_1	0.112	0.209	0.547	0.543	0.330
面板 se	(0.096)	(0.374)	(0.327)	(0.209)	(0.110)
异方差 se	[0.079]	[0.423]	[—]	[0.226]	[—]
默认 se	{0.023}	{0.389}	{—}	{0.169}	{—}
RMSE	0.283	0.296	0.307	0.307	0.298
工具	5	9	9	72	72
OIR 检验	—	—	5.45	—	69.51
dof	—	—	4	—	67
p 值	—	—	0.244	—	0.393
N	4 256	4 256	4 256	4 256	4 256

[a] 差分回归使用了 1981～1988 年期间 523 人的年度数据。报告的是 β_1、$\Delta\ln wg$ 的系数,以及三种估计标准误差:圆括号中数值为面板稳健的,方括号中数值为异方差稳健的,而假定误差的默认估计值在大括号中。另外,所有回归元包括 Δkids、Δage、Δagesq 以及 Δdisab,却没有报告它们的系数估计。工具是滞后两时期的 lnwg、kids、age 以及既有滞后一个时期又有滞后两时期的 disab。对于基准情况,存在 9 个工具,而对于叠放情况,存在 8×9=72 个工具。RMSE 表示残差的均方误差平方根,OIR 表示过度识别约束检验统计量,dof 表示自由度,而 p 值表示检验的 p 值。

OLS: OLS 列报告了式(22.22)的 OLS 估计。其劳动供给弹性 0.12 稍微不同于表 21.2 中一阶差分列中的估计值 0.109,因为那里还包括四个人口统计变量作为回归元,而且省略了另外一年的数据。由于一阶差分进行建模,其模型拟合表现差,而包括截距的 R^2 是 0.006。

基准情况工具的 2SLS: 基准情况中的工具使用由式(22.11)定义的 $\mathbf{Z}_i$,其中,$\mathbf{z}_{it}$ 拥有 9 个元素: $\ln wg_{i,t-2}$, $kids_{i,t-1}$, $age_{i,t-1}$, $agesq_{i,t-1}$, $disab_{i,t-1}$, $kids_{i,t-2}$, $age_{i,t-2}$, $agesq_{i,t-2}$ 以及 $disab_{i,t-2}$。于是,此模型的 9 个工具是过度识别的,而 5 个参数是待估的。β_1 的 2SLS 估计值的准确性比 OLS 估计值的要差一些,其标准误差从 0.096 增大到 4 倍的 0.374。对于其他回归元则没有报告,其有效性损失也不小。

叠放工具的 2SLS: 基准情况是建立在 9 个矩条件 $\mathrm{E}[\sum_{t=3}^{10}\mathbf{z}_{it}u_{it}]=\mathbf{0}$ 基础上的 GMM。相反,叠放工具使用 72(=8×9)个矩条件 $\mathrm{E}[\mathbf{z}_{it}u_{it}]=\mathbf{0}$, $t=3,\cdots,10$,其中,$\mathbf{z}_{it}$ 如同基准情况一样。于是,使用由式(22.14)定义的 $\mathbf{Z}_i$,这里,$\mathbf{Z}_i$ 表示 8 年 72 个工具。$\mathbf{Z}_i$ 的第 i 行是由式(22.17)给出的,此处,$\mathbf{z}_{it}$ 表示基准情况工具的 9×1 列向量。为了建立工具,首先生成对于所有 i 与 t 的 72 个变量。变量 ztj 等于 0,其中,t 表示年份,而 j 表示第 j 个工具。然后,当 $t=s$ 时,就用 $z_{it,j}$ 代替 zsj_{it};而当

$t \neq s$ 时，则令 $zsj_{it}=0$。例如，当第五个工具为 $\text{disab}_{i,t-1}$时，如果 $t=3$(第三年度)，那么令 $z35$ 等于 $\text{disab}_{i,2}$；而对于 $t\neq3$，令等于 $zt5$ 等于 0。从而，2SLS 估计值能通过 Δlnhrs_{it} 对式(22.22)中 5 个回归元进行标准 2SLS 回归并将这 72 个构造变量作为工具来获得。一旦利用扩展工具，就得到 2SLS 估计的标准误差从 0.374 降到 0.209，而且是最初 OLS 估计值的 2 倍。

两步 GMM： 表 22.2 中的两步 GMM 估计值不同于齐利亚克(Ziliak，1997)表 1 里的那些估计值，因为此处由式(22.7)定义的 $\hat{\mathbf{S}}$ 的面板稳健估计值用于建立加权矩阵，而齐利亚克则使用异方差稳健 $\hat{\mathbf{S}}=N^{-1}\sum_i\hat{u}_{it}^2\mathbf{z}_{it}\mathbf{z}_{it}'$。正如人们所料，两步 GMM 估计量比 2SLS 更加有效，它的标准误差从含有基准工具的 0.374 降到 0.327，并从含有叠放工具的 0.209 降到 0.110。最后这个标准误差并不比 OLS 的大多少。

过度识别约束检验： 关于这度识别约束的检验统计量已由式(22.10)给出。由表 22.2 知，基准情况和叠放工具的检验统计量均具有比 0.05 更大的 p 值，因此，并没有拒绝约束，我们得出结论，过度识别工具都是有效工具。

弱工具检验： 对弱工具的诊断已经在 22.2.4 节与 5.9 节讨论过。由于没有一个回归元出现在工具集合中，所以要使用源于第一阶段回归的整个 F 统计量，而不是回归元子集 F 统计量。对于基准情况工具而言，Δlnwg 对 9 个工具与常数项进行回归得出，面板稳健的 $F=2.80$，类似地对 72 个叠放工具进行回归，得出 $F=1.90$，这表明有限样偏倚极有可能出现。对于 Δkids、Δage、Δagesq 以及 Δdisab，式(22.22)中一些回归元同样被处理成内生的进行类似回归，得出所有情况下 $F>8.5$，关于 Δlnwg 的谢伊偏 R^2(参见 4.9.4 节)是 0.003 6，大于其他 4 个内生回归元的 0.075。因此，弱工具问题归因于寻找 Δlnwg 的好工具问题。

有效性提高： 在此例中，面板 GMM 估计量被用于控制内生性。然而，即使假定所有回归元均是强外生的，面板 GMM 仍是引人注目的，因为它比 OLS 更有效，除非误差是 iid 的；参见式(22.20)后面的讨论。举一个例子，含有工具的面板两步 GMM 估计量设置基准情况工具以及式(22.22)中的 5 个最初回归元，得出 $\hat{\beta}_1=0.016$，其标准误差为 0.076，小于 OLS 标准误差 0.096。

22.4 随机效应与固定效应面板 GMM

现在，我们通过包括时常值可加特定个体效应(**individual-specific effect**)α_i，扩大面板数据模型(22.1)，因而有：

$$y_{it}=\alpha_i+\mathbf{x}_{it}'\boldsymbol{\beta}+\varepsilon_{it} \tag{22.23}$$

于是，式(22.1)的误差项现在建模成为 $u_{it}=\alpha_i+\varepsilon_{it}$。为了简单起见，同样记号既用于固定效应又用于随机效应模型，如同随机效应模型一样的情况，21.7 节的共同截距 μ 被归入 $\mathbf{x}_{it}'\boldsymbol{\beta}$之中。

假定回归元 $\mathbf{x}_{it}$的一些分量是内生的，满足 $\mathrm{E}[\mathbf{x}_{it}(\alpha_i+\varepsilon_{it})]\neq\mathbf{0}$，所以 $\boldsymbol{\beta}$ 的估计量是非一致的。本节在各种背景下，包括固定效应、随机效应、固定效应与随机效

应的混合以及联立方程，提出获得 $\boldsymbol{\beta}$ 的一致估计值的 IV 估计量。

22.4.1 *是随机效应还是固定效应？*

回顾第 21 章，特定个体效应 α_i 既能在 FE 模型中又能在 RE 模型中被处理成随机的。这个随机变量 α_i 与 $\mathbf{x}_{it}$ 在 RE 模型中是独立的，但在 FE 模型中 α_i 与 $\mathbf{x}_{it}$ 却是相关的。对于 RE 模型，所有系数都是可估计的，而在 FE 模型中，时常值回归元的系数却不是可估计的，因为一致估计需要通过差分去掉 α_i 与时常值回归元。

在本章含有内生回归元的情况下，我们认为，模型是随机效应模型，如果工具 $\mathbf{Z}_i$ 存在，满足 $\mathrm{E}[\mathbf{Z}_i'(\alpha_i+\varepsilon_{it})]=\mathbf{0}$。于是，22.2 节的方法将对所有回归参数进行一致估计成为可行。相反，如果寻找一些工具，使得 $\mathrm{E}[\mathbf{Z}_i'\varepsilon_{it}]=\mathbf{0}$ 但 $\mathrm{E}[\mathbf{Z}_i'\alpha_i]\neq\mathbf{0}$ 是可能的，我们就认为此模型是固定效应模型。于是，必须通过进行差分去掉 α_i，在此情况下，仅有时变回归元的系数将是可识别的。

22.4.2 *固定效应模型 IV*

若将 21.2 节给出的各种不同差分运算应用到式(22.23)，则得到变换模型(**transformed model**)形式：

$$\tilde{y}_{it}=\tilde{\mathbf{x}}_{it}'\boldsymbol{\beta}+\tilde{\varepsilon}_{it}$$

其中，“～”表示通过差分变换去掉 α_i 的符号，而一些重要例子将由下面给出。一旦进行叠放，我们得出：

$$\tilde{\mathbf{y}}_i=\widetilde{\mathbf{X}}_i\boldsymbol{\beta}+\tilde{\boldsymbol{\varepsilon}}_i \tag{22.24}$$

如果 $\mathrm{E}[\mathbf{x}_{it}\varepsilon_{it}]\neq 0$，那么 $\mathrm{E}[\tilde{\mathbf{x}}_{it}\tilde{\varepsilon}_{it}]\neq 0$，而对式(22.24)的 LS 估计得到非一致估计。

现在倘若工具 $\mathbf{Z}_i$ 存在，满足 $\mathrm{E}[\mathbf{Z}_i'\tilde{\boldsymbol{\varepsilon}}_i]\neq\mathbf{0}$，我们考察 IV 估计，于是，式(22.24)具有工具 $\mathbf{Z}_i$ 的面板 GMM 估计(IV、2SLS 或者 2SGMM)，得出时变回归元系数的一致估计。

获得工具的一种方式是，通过类似于横截面情况的推理方法来进行。有效工具是与回归元相关但不与误差相关的变量，但也可通过从式(22.23)右边排除来进行。另一种获得工具的方式是这里特别强调的方式，若利用 22.2.4 节详述的外生性假设，运用不是当前时期的一些时期外生回归元。

关于工具可用性的原始假设是那些 $\mathbf{z}_{is}$ 与 ε_{it} 之间的相关假设。然而，此处它是 $\mathbf{z}_{is}$ 与起作用的差分误差 $\tilde{\varepsilon}_{it}$ 之间的相关假设。通常，必须剔除固定效应，进行差分减少可利用工具的数目。一些差分运算导致的损失比另一些差分运算要大一些，并且甚至能产生非一致的估计。我们考察关注弱外生工具的三种不同差分运算。在实际应用中，尤其是针对动态模型的应用，这是一种更现实的假设。

一阶差分 IV 模型

一阶差分 IV 估计量是关于一阶差分模型：

$$y_{it}-y_{i,t-1}=(\mathbf{x}_{it}-\mathbf{x}_{i,t-1})'\boldsymbol{\beta}+(\varepsilon_{it}-\varepsilon_{i,t-1}),\quad t=2,\cdots,T \tag{22.25}$$

的 IV 或 2SLS 或面板估计量。其弱外生性假设：$\mathrm{E}[\mathbf{z}_{is}\varepsilon_{it}]=\mathbf{0}$，对于 $s\leqslant t$，蕴含着

$E[\mathbf{z}_{is}(\varepsilon_{it}-\varepsilon_{i,t-1})]=\mathbf{0}$，对于 $s\leqslant t-1$。因此，一阶差分使得可利用工具集合的时间序列缩短了一个时期，所以仅有 $\mathbf{z}_{i,t-1},\mathbf{z}_{i,t-2},\cdots$ 可作为工具。当假定弱外生性，就会得到一致估计量。

使用滞后回归元作为工具，首先是由安德森和萧政（Anderson and Hsiao，1981）在动态面板模型背景下提出的，而后由霍尔茨·埃金、纽韦和罗森（Holtz-Eakin, Newey, and Rosen, 1988）以及阿雷拉诺和邦德（Arellano and Bond, 1991）（参见 22.5.3 节）加以推广。

注意到，人们能使用变换工具 $\tilde{\mathbf{z}}_{is}=\Delta\mathbf{z}_{is}=\mathbf{z}_{is}-\mathbf{z}_{i,s-1}$，$s\leqslant t-1$。然而，这样做并不存在什么好处，因为利用 $\Delta\mathbf{z}_{i,t-1},\cdots,\Delta z_{i2},\mathbf{z}_{i1}$ 等价于利用 $\mathbf{z}_{i,t-1},\cdots,\mathbf{z}_{i2},\mathbf{z}_{i1}$ 作为工具，并且如果数据以第 1 个时期开始，那么仅能计算出 $\mathbf{z}_{i1}$，而不能计算 $\Delta\mathbf{z}_{i1}$。

组内或均值差分 IV 模型

组内估计量是关于组内模型或均值差分模型：

$$y_{it}-\bar{y}_i=(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)'\boldsymbol{\beta}+(\varepsilon_{it}-\bar{\varepsilon}_i) \tag{22.26}$$

的 IV 或 2SLS 或面板 GMM 估计量。于是，$E[\mathbf{z}_{is}\varepsilon_{it}]=\mathbf{0}$，对于 $s\leqslant t$，不再蕴含着 $E[\mathbf{z}_{is}(\varepsilon_{it}-\bar{\varepsilon}_i)]=\mathbf{0}$，甚至对于比 t 很小的 s。为了理解这一点，假定 $E[\mathbf{z}_{is}\varepsilon_{it}]\neq\mathbf{0}$，对于 $s>t$。于是，$E[\mathbf{z}_{it}\bar{\varepsilon}_i]\neq\mathbf{0}$，对于所有 s，因为 $\bar{\varepsilon}_i=T^{-1}\sum_t\varepsilon_{it}$ 包括了过去 ε_{it}，这与 $\mathbf{z}_{is}$ 是相关的。

因而，若工具是弱外生的，或若工具满足甚至同时期外生性的较弱假设或求和条件，则组内模型的 IV 估计导致 $\boldsymbol{\beta}$ 的非一致估计。如果工具确定是强外生的，那么只能使用组内变换。

向前正交推导 IV 模型

对一阶差分的一种可供选择方法是，同样需要工具只是弱外生的而不是强外生的，此方法是由阿雷拉诺和博韦（Arellano and Bover, 1995）提出。尽管人们已经广泛使用一阶差分，但我们还是要阐述该方法。

对于第 i 个观测值的叠放模型（22.2），一阶差分变换得出模型 $\mathbf{Dy}_i=\mathbf{DX}_i\boldsymbol{\beta}+\mathbf{D}\boldsymbol{\varepsilon}_i$，其中，$\mathbf{D}$ 表示 $(T-1)\times T$ 阶矩阵，其元素为 $\mathbf{D}_{ts}$，$t=1,\cdots,T-1$，$s=1,\cdots,T$，当 $s=t$ 时，$\mathbf{D}_{ts}=-1$，当 $s=t+1$ 时，$\mathbf{D}_{ts}=1$，否则 $\mathbf{D}_{ts}=0$。若 ε_{it} 是 iid 的，则变换误差是 MA(1)的且 $V[\mathbf{Du}_i]=\sigma^2\mathbf{DD}'$。于是，GLS 估计量利用 $(\mathbf{DD}')^{-1/2}$ 左乘 $\mathbf{D}\boldsymbol{\varepsilon}_i$，或者利用 $(\mathbf{DD}')^{-1/2}\mathbf{D}$ 左乘 $\boldsymbol{\varepsilon}_i$，从而得到变换模型形式（22.24），其中，"～"表示利用 $(\mathbf{DD}')^{-1/2}\mathbf{D}$ 左乘的形式。

如果使用上三角乔列斯基（Cholesky）因子分解法获得 $(\mathbf{DD}')^{-1/2}$，这就得出向前正交推演模型（**forward orthogonal deviation model**）：

$$c_t(y_{it}-\bar{y}_{it}^F)=c_t(\mathbf{x}_{it}-\bar{\mathbf{x}}_{it}^F)'\boldsymbol{\beta}+c_t(\varepsilon_{it}-\bar{\varepsilon}_{it}^F) \tag{22.27}$$

[参见阿雷拉诺（Arellano, 2003，第 17 页）。]其中，$c_t^2=(T-t)/(T-t+1)$，而上标"F"表示仅仅使用未来值用于求平均值。例如，$\bar{y}_{it}^F=(T-1)^{-1}\sum_{s=t+1}^{T}y_{is}$。

此变换称为正交推导（**orthogonal deviation**），因为变换误差 $c_i(\varepsilon_{it}-\bar{\varepsilon}_i^F)$ 具有单位方差且是无关的。添加形容词"向前"表示变换误差只依赖于最初误差的当前值与未来值。对式（22.27）进行 OLS 估计得到第 21 章的组内估计量，因此，若实际

上 ε_i 是 iid 的，则正交推导变换是最优的。

向前正交推导估计量（**forward orthogonal deviation IV estimator**）是模型(22.27)的 IV 或 2SLS 或面板 GMM 估计量。对于弱外生工具，当 $s \leqslant t$，$\mathrm{E}[\mathbf{z}_{is}\varepsilon_{it}]=\mathbf{0}$ 蕴含着 $\mathrm{E}[\mathbf{z}_{is}(\varepsilon_{it}-\varepsilon_i^F)]=\mathbf{0}$。因此，向前正交推导并不会导致可利用工具数目的损失。通常，此变换不能应用于工具，因为$(\mathbf{z}_{it}-\mathbf{z}_i^F)$ 涉及 $\mathbf{z}_{it}$ 未来值，在许多应用中 $\mathbf{z}_{it}$ 与 ε_{it} 是相关的。

22.4.3 *随机效应模型 IV*

关于第 i 个观测值的叠放模型是：

$$\mathbf{y}_i=\mathbf{X}_i\boldsymbol{\beta}+\mathbf{e}\alpha_i+\boldsymbol{\varepsilon}_i$$

其中，$\mathbf{e}$ 表示 $T\times 1$ 维单位向量。给定工具 $\mathbf{Z}_i$ 时，通过直接应用 22.2 节的面板 GMM 估计量获得的一致但无效的估计值，通过排除性约束或通过合适外生性约束来得到，使得 $\mathrm{E}[\mathbf{Z}_i'(\mathbf{e}\alpha_i+\boldsymbol{\varepsilon}_i)]=\mathbf{0}$。这里，我们进一步探讨并考察更有效的估计，如同第 21 章一样，控制给定误差成分模型 $u_{it}=\alpha_i+\varepsilon_{it}$ 时不同时间上的误差相关性。

变换模型 IV 估计

假定工具 $\mathbf{Z}_i$ 满足 $\mathrm{E}[\mathbf{u}_i|\mathbf{Z}_i]=\mathbf{0}$ 且 $\mathrm{V}[\mathbf{u}_i|\mathbf{Z}_i]=\boldsymbol{\Omega}_i$，其中，$\boldsymbol{\Omega}_i$ 具有与标准模型相同的形式，它的对角元素为 $\sigma_\alpha^2+\sigma_\varepsilon^2$，而非对角元素为 σ_α^2。注意，这是比 $\mathrm{E}[\mathbf{Z}_i'\mathbf{u}_i]=\mathbf{0}$ 更强的假设，从而对利用工具施加了约束。

给定条件矩条件 $\mathrm{E}[\mathbf{u}_i|\mathbf{Z}_i]=\mathbf{0}$，由 6.3.7 节知，最优无条件矩条件是：

$$\mathrm{E}[\mathbf{Z}_i'\boldsymbol{\Omega}_i^{-1}\mathbf{u}_i]=\mathrm{E}[(\boldsymbol{\Omega}_i^{-1/2}\mathbf{Z}_i)'(\boldsymbol{\Omega}_i^{-1/2}\mathbf{u}_i)]=\mathbf{0}$$

这就导致了对含有变换工具 $\mathbf{Z}_i^*$ 的变换方程组 $\mathbf{y}_i^*=\mathbf{X}_i^*\boldsymbol{\beta}+\mathbf{u}_i^*$ 的 GMM 估计，其中，"$*$"表示利用 $T\times T$ 阶矩阵 $\boldsymbol{\Omega}_i^{-1/2}$或一致估计值 $\hat{\boldsymbol{\Omega}}_i^{-1/2}$ 左乘。

由 21.7.1 节知，左乘 $\hat{\boldsymbol{\Omega}}_i^{-1/2}$ 会得到下面模型：

$$y_{it}-\hat{\lambda}\bar{y}_i=(\mathbf{x}_{it}-\hat{\lambda}\,\bar{\mathbf{x}}_i)'\boldsymbol{\beta}+\{(1-\hat{\lambda})\alpha_i+(\varepsilon_{it}-\hat{\lambda}\bar{\varepsilon}_i)\} \qquad (22.28)$$

其中，$\hat{\lambda}$ 表示$\lambda=1-\sigma_\varepsilon/\sqrt{\sigma_\varepsilon^2+T\sigma_\alpha^2}$的一致估计值。随机效应 IV 估计量是具有变换工具 $\tilde{\mathbf{z}}_{it}=(\mathbf{z}_{it}-\hat{\lambda}\bar{\mathbf{z}}_i)$或等价地具有工具 $\mathbf{z}_{it}-\bar{\mathbf{z}}_i$ 与 $\bar{\mathbf{z}}_i$ 的模型的 IV 或 2SLS 估计量。

此模型需要 λ 的一致估计值$\hat{\lambda}$。对于 σ_ε^2，我们使用 $\hat{\sigma}_\varepsilon^2=\sum_{it}\hat{\varepsilon}_{it}^2/N(T-1)$，其中，$\hat{\varepsilon}_{it}$表示源于组内回归的残差，其工具为$(\mathbf{z}_{it}-\bar{\mathbf{z}}_i)$[参见式(22.26)]。同样地，$\sigma_\varepsilon^2+T\sigma_\alpha^2$ 能通过$\sum_i\hat{\bar{u}}_i^2/N$ 加以估计，其中，$\hat{\bar{u}}_i$ 表示源于 $\bar{y}_i$ 对$\bar{\mathbf{x}}_i$ 进行组内回归的残差，其工具为 $\bar{\mathbf{z}}_i$。所得到的估计量被巴尔塔吉(Baltagi，1981)称为误差成分估计量(**error components 2SLS estimator**，记为 EC2SLS)。

这些模型均依赖于对 $\boldsymbol{\Omega}_i$ 所设定的特殊函数形式。一旦利用式(22.5)，其中，$\mathbf{y}$、$\mathbf{X}$、$\mathbf{Z}$ 以及 $\mathbf{W}_N=[\mathbf{Z}'\mathbf{Z}]^{-1}$都要用式(22.28)中的变换变量代替，22.2.2 节的结果使得对错误设定来说是稳健的估计成为可能。

一个更重要的约束是，只有最初工具是强外生的，才能使用此方法。这里，一致

性需要 $E[\mathbf{Z}_i'\mathbf{\Omega}_i^{-1}\mathbf{u}_i]=\mathbf{0}$，即比 $E[\mathbf{Z}_i'\mathbf{u}_i]=\mathbf{0}$ 更强一些的假设，实际上需要 $E[\mathbf{u}_i|\mathbf{Z}_i]=\mathbf{0}$。例如，假定 $E[\mathbf{z}_{it}\alpha_i]=\mathbf{0}$，对于所有 t，然而 $E[\mathbf{z}_{it}\varepsilon_{it}]=\mathbf{0}$ 对于 $s\leqslant t$，但 $E[\mathbf{z}_{it}\varepsilon_{it}]\neq\mathbf{0}$，对于 $s>t$。于是，$E[\mathbf{z}_{it}\bar{\varepsilon}_i]\neq\mathbf{0}$，导致了式(22.28)的含有误差项工具的相关性。

22.4.4 *豪斯曼—泰勒混合模型 IV*

内生性的一个重要例子涉及回归元与特定个体效应 α_i 是相关的。这导致了第 21 章估计量的非一致性。一种明显求解法反而使用组内(或固定效应)估计量，它是一致的。然而，时常值个体回归元的系数却是不能识别的。这使得许多面板研究的目的即估计时常值回归元效应受挫，诸如后学校教育工资回归中受教育程度的效应。

豪斯曼和泰勒(Hausman and Taylor, 1981)考察了下式对式(22.23)的变形：

$$y_{it}=\mathbf{x}_{1it}'\boldsymbol{\beta}_1+\mathbf{x}_{2it}'\boldsymbol{\beta}_2+\mathbf{w}_{1i}'\boldsymbol{\gamma}_1+\mathbf{w}_{2i}'\boldsymbol{\gamma}_2+\alpha_i+\varepsilon_{it} \tag{22.29}$$

其中，假定一些回归元与 α_i 是相关的，而另一些回归元与 α_i 则是不相关的，同时引进 $\mathbf{w}$ 表示时常值回归。特别地，$\mathbf{x}_{1it}$ 及 $\mathbf{w}_{1i}$ 与 α_i 是不相关的，但 $\mathbf{x}_{2it}$ 及 $\mathbf{w}_{2i}$ 与 α_i 是相关的。假定所有回归元与 ε_{it} 是不相关的。在此模型中，α_i 被认为是随机效应和固定效应的混合(**hybrid**)。

豪斯曼和泰勒(Hausman and Taylor, 1981)提出以两种方式利用时变外生回归元 $\mathbf{x}_{1it}$：为估计 $\boldsymbol{\beta}_1$ 与作为 $\mathbf{w}_{2i}$ 的工具，使得对 $\boldsymbol{\gamma}$ 的估计可行。于是，如果时变外生回归元的数目等于或大于时常值内生回归元的数目，那么 $\boldsymbol{\gamma}$ 是可识别的。雨宫和麦柯迪(Amemiya and MaCurdy, 1986)提出以$(T+1)$种方式使用 $\mathbf{x}_{1it}$ 的更有效估计量：为估计 $\boldsymbol{\beta}_1$ 与作为 $\mathbf{w}_{2i}$ 的工具，当 $\dim[\mathbf{w}_{2i}]\geqslant T\dim[\mathbf{x}_{1it}]$ 时，则可以识别。利用非当前时期的其他时期外生回归元作为工具的方法，已经在 22.2.4 节详细讨论了。

各种投影中有一些是等价的，能用于生成合适的工具。布鲁什、米宗和施密特(Breusch, Mizon, and Schmidt, 1989)提供了允许利用 2SLS 软件包进行估计的较简单阐述与投影。

首先，考察忽略$(\alpha_i+\varepsilon_{it})$相关结构的一致却无效的估计。组内变换剔除了与 α_i 的相关，因此，$\ddot{\mathbf{x}}_{2it}=\mathbf{x}_{2it}-\bar{\mathbf{x}}_{2i}$ 能用作内生 $\mathbf{x}_{2it}$ 的工具。类似地，$\mathbf{x}_{1it}$ 的工具是 $\ddot{\mathbf{x}}_{1it}$，而不是更明显的 $\mathbf{x}_{1it}$。于是，$\bar{\mathbf{x}}_{1i}$ 用作内生 $\mathbf{w}_{2i}$ 的工具，而内生 $\mathbf{w}_{1i}$ 用作它自身的工具。

现在，在随机效应假设即分量 α_i 与 ε_{it} 都是同方差的假设下，考察有效估计问题。于是，由式(22.27)知，随机效应差分变换[参见式(22.28)]导致：

$$\tilde{y}_{it}=\tilde{\mathbf{x}}_{1it}'\boldsymbol{\beta}_1+\tilde{\mathbf{x}}_{2it}'\boldsymbol{\beta}_2+\tilde{\mathbf{w}}_{1i}'\boldsymbol{\gamma}_1+\tilde{\mathbf{w}}_{2i}'\boldsymbol{\gamma}_2+v_{it} \tag{22.30}$$

例如，这里 $\tilde{\mathbf{x}}_{1it}=\mathbf{x}_{1it}-\hat{\lambda}\,\bar{\mathbf{x}}_{1i}$，其中，纯量 $\hat{\lambda}$ 的估计量已在上一节末尾阐述过。豪斯曼—泰勒估计量等价于利用工具 $\ddot{\mathbf{x}}_{1it}$，$\ddot{\mathbf{x}}_{2it}$，$\mathbf{w}_{1i}$ 以及 $\bar{\mathbf{x}}_{1i}$ 的式(22.30)IV 估计。外生时变回归元 $\mathbf{x}_{1it}=\ddot{\mathbf{x}}_{1it}+\bar{\mathbf{x}}_{1i}$ 用作两次工具，即组内差分 $\ddot{\mathbf{x}}_{1it}$ 用于 $\mathbf{x}_{1it}$ 的工具，而时间均值用于 $\mathbf{w}_{2i}$ 的工具。雨宫和麦柯迪(Amemiya and MaCurdy, 1986)反而使用工具 $\ddot{\mathbf{x}}_{1it}$、$\ddot{\mathbf{x}}_{2it}$、$\ddot{\mathbf{w}}_{1it}$ 和 $\ddot{\mathbf{x}}_{1it}$，…，$\ddot{\mathbf{x}}_{1iT}$，因而是把 $\mathbf{x}_{1i}$ 的全部历史而非时间均值用作工具。这需

要对于 $t=1,\cdots,T$，有比 $\mathrm{E}[\bar{\mathbf{x}}_{1i}\alpha_i]=\mathbf{0}$ 更强的假设（参见 22.2.4 节）。布鲁什等人（Breusch et al.，1989）提出利用 $\ddot{\mathbf{x}}_{2is}$ 作为额外工具的甚至更有效的估计量，对于 $s\neq t$ 。

这种方法的主要局限性是，它需要设定一些回归元与 α_i 相关，或者不相关。在后受教育对数工资回归中，豪斯曼和泰勒以做出下述假定开始：假定所有三个时变回归元（经历、差的健康状况、去年失业）都是外生的，两个时常值回归元（种族与联合工会）是外生的，而关注的时常值回归元（受教育）是内生的。这种设定存在两个过度识别约束。模型设定检验可能要利用基于 $\hat{\beta}_{\mathrm{HT}}$ 与 $\hat{\beta}_{\mathrm{W}}$ 之差的豪斯曼检验，不论 $\mathbf{x}_{it}$ 与 $\mathbf{w}_i$ 的哪些分量与 α_i 相关，因为关于 β 的组内估计量是一致的。康沃尔和鲁珀特（Cornwall and Rupert，1988）提供了对比各种估计量的实证研究。

22.4.5 *SUR* 与联立方程估计

以上的面板数据分析，以独立方式全部地关注单方程估计。在一些情况下，人们希望估计方程组，诸如需求方程组，其中，因变量与回归元对许多个体来说在一些时点均是可观测的。如果参数不存在交叉方程约束，那么单方程估计会产生一致估计，但利用联合方程估计可能获得更有效的估计，这里的联合方程估计运用了不同方程的误差相关。

在第 21 章强外生回归元框架下，更有效估计量是看似不相关回归从横截面到面板数据的推广。误差成分 SUR 模型（**error components SUR model**）对 G 个方程的第 g 个方程设定如下：

$$y_{git}=\mathbf{x}'_{git}\boldsymbol{\beta}+\alpha_{gi}+\varepsilon_{git},\quad g=1,\cdots,G \tag{22.31}$$

如同横截面情况一样，α_{gi} 对于不同 i 是独立的，ε_{git} 对于不同 i 与 t 是独立的，而且 α_{gi} 与 ε_{git} 是相互独立的。然而，允许误差成分对不同成分是相关的，因此 $\mathrm{Cov}[\alpha_{gi},\alpha_{hi}]\neq 0$，且 $\mathrm{Cov}[\varepsilon_{git},\varepsilon_{hit}]\neq 0$，对于 $g\neq h$。于是，第 21 章的一些方法会产生一致估计值。显而易见的单方程估计量是随机效应估计量，它是控制组内每一个方程相关的可行 GLS。额外控制误差交叉方程相关的更有效估计量已由埃弗里（Avery，1977）和巴尔塔吉（Baltagi，1980）详细阐述过。

当方程组是联立方程系统时，能建立类似的有效性提高，其中，式（22.31）中回归元 $\mathbf{x}_{git}$ 现在可能包括来自其他方程的一个或多个内生回归元 y_{hit}。于是，对每个单方程进行 IV 或 GMM 估计会得出一致估计值，显而易见，该估计量给出了误差成分结构是 22.4.3 节的随机效应 IV 或 EC2SLS 估计量。一旦利用由巴尔塔吉（Baltagi，1981）提出的误差成分三阶段最小二乘法估计量，就可通过系统估计来获得更有效的估计值。

系统估计量更加难以实施，而对每个方程分别进行估计可能是恰当的。不过，即使采用这一较简单方法，在设定联立方程组时，大部分都可以获得，因为它允许利用从关注方程中排除的作为工具的外生回归元。这提供了比利用来自非当前时期的其他时期外生回归元作为工具情况更为传统的获得工具的方法。

22.5 动态模型

在本节,我们考察通常的特定个体效应面板数据模型,其复杂情况就是回归元包括滞后因变量。于是,模型满足:

$$y_{it}=\gamma y_{i,t-1}+\mathbf{x}_{it}'\boldsymbol{\beta}+\alpha_i+\varepsilon_{it},\quad i=1,\cdots,N,\ t=2,\cdots,T^{[1]} \tag{22.32}$$

与以往一样,面板数据是短的,并对于不同 i 是独立的。假定 $|\gamma|<1$, 即 22.5.4 节所放松的假设。

一个重要结果是,尽管 α_i 是随机效应,但式(22.32)的 OLS 估计会产生 γ 与 $\boldsymbol{\beta}$ 的非一致估计。这是因为回归元 $y_{i,t-1}$ 与 α_i 相关,从而与综合误差($\alpha_i+\varepsilon_{it}$)相关。甚至对于随机效应,仍需要一种可供选择的估计量。

当 α_i 是固定效应、$|\gamma|<1$、误差 ε_{it} 是序列无关的,并且面板是短的(参见 22.5.3 节)时,我们考察其估计问题。虽然这是微观经济计量学应用的基本情况,但仍存在大量文献致力于对这些假设中的一个或多个加以改进。更一般地,特定个体效应可能是纯随机的,误差可能是序列相关的,数据可能是非平稳的,而面板也可能是长面板数据,但我们几乎没有谈及此类文献。

22.5.1 *真实状态相依性与不可观测异质性*

在考察估计之前,我们注意到,y_{it} 的时间序列相关除由第 21 章考虑的经由 α_i 的间接效应引起之外,现今直接由 $y_{i,t-1}$ 而引起。这两种原因导致了例如个体收入或接受福利救济对于不同时间相关的截然不同的解释。

为了简单起见,设 $\boldsymbol{\beta}=\mathbf{0}$,所以 $y_{it}=\gamma y_{i,t-1}+\alpha_i+\varepsilon_{it}$。于是,$\mathrm{E}[y_{it}|y_{i,t-1},\alpha_i]=\gamma y_{i,t-1}+\alpha_i$,并且 $\mathrm{Cor}[y_{it},y_{i,t-1}|\alpha_i]=\gamma$。对于仅由自相关参数 γ 决定的 y_{it} 方面的不同时间相依性来说,以 α_i 为条件,关于 AR(1)模型的标准时间序列结果可以应用。然而,α_i 是未知的,我们实际上观测到 $\mathrm{E}[y_{it}|y_{i,t-1}]=\gamma y_{i,t-1}+\mathrm{E}[\alpha_i|y_{i,t-1}]$ 并且 $\mathrm{Cor}[y_{it},y_{i,t-1}]\neq\gamma$。特别地,由满足 $\boldsymbol{\beta}=\mathbf{0}$ 的式(22.32)知:

$$\begin{aligned}\mathrm{Cor}[y_{it},y_{i,t-1}]&=\mathrm{Cor}[\gamma y_{i,t-1}+\alpha_i+\varepsilon_{it},y_{i,t-1}]\\&=\gamma+\mathrm{Cor}[\alpha_i,y_{i,t-1}]\\&=\gamma+\frac{(1-\gamma)}{1+(1-\gamma)\sigma_\varepsilon^2/(1+\gamma)\sigma_\alpha^2}\end{aligned} \tag{22.33}$$

其中,第二个等式假定 $\mathrm{Cor}[\varepsilon_{it},y_{i,t-1}]=0$,而第三个等式在经过关于随机效应含有 ε_{it} iid $[0,\sigma_\varepsilon^2]$ 且 α_i iid $[0,\sigma_\alpha^2]$ 的特殊情况的一些代数运算之后,就可以获得。

结果式(22.33)清楚表明,y_{it} 与 $y_{i,t-1}$ 之间相关存在两种可能原因。

当不同时间相关是因为 $y_{i,t-1}$ 上一个时期决定 y_{it} 这个时期的因果机制,就出现了真实状态相依性(**true state dependence**)。如果个体效应 $\alpha_i\simeq0$,从而 $\mathrm{Cor}[y_{it},$

[1] 原著中该式 t 取值为"$t=1,\cdots,T$",依据上下文判断,可能是一个印刷错误,应该为"$t=2,\cdots,T$"。——译者注

$y_{i,t-1}]\simeq\gamma$,那么这种相依性是相对大的。更一般地,相对于 σ_ε^2,σ_α^2 是非常小的,就出现此情况。

即使不存在因果机制,归因于不可观测异质性的相关性出现,因此 $\gamma=0$,但当 $\gamma=0$ 时,由于 $\mathrm{Cor}[y_{it},y_{i,t-1}]$ 简化成 $\sigma_\alpha^2/(\sigma_\alpha^2+\sigma_\varepsilon^2)$,所以存在相关,如同第 21 章一样。

两种极端允许这种相关性任意地接近于 1,因为它们的 $\gamma\to 1$ 或 $\sigma_\varepsilon^2/\sigma_\alpha^2\to 0$。不过,这些针对相当不同的政策含义,给出了两种截然不同的解释。关于在控制回归元 x_{it} 之后,收入 y_{it} 作为随时间推移连续高的真实状态相依性解释是,未来收入是由过去收入决定的,同时很大。而不可观测异质性解释是,实际上 γ 很小,但重要变量已经从中 x_{it} 省略,从而导致每个时期中的 α_i 很大。对于持续期间限数据来说,真实状态相依性与不可观测异质性之间的区别已在第 18 章探讨过。第 21 章的静态线性面板模型仅仅考虑了不可观测异质性。

22.5.2 *标准面板估计量的非一致性*

如果回归元包括滞后因变量,甚至在随机效应模型情况下,那么来自上一章的一些估计量都是非一致的。我们考察由式(22.32)给出的模型估计,其中的文献通常假定 ε_{it} 是序列无关的。

首先,考察 y_{it} 对 $y_{i,t-1}$ 与 $\mathbf{x}_{it}$ 的 OLS 估计。于是,误差项是($\alpha_i+\varepsilon_{it}$),这与回归元 $y_{i,t-1}$ 相关,因为滞后方程为 $y_{i,t-1}=\gamma y_{i,t-2}+\mathbf{x}_{i,t-1}'\boldsymbol{\beta}+\alpha_i+\varepsilon_{i,t-1}$,因而 $y_{i,t-1}$ 与 α_i 相关。注意到,这违背了前面不带滞后因变量的随机效应模型 OLS 估计的结果,从而 y_{it} 对 $\mathbf{x}_{it}$ 的 OLS 会得出虽然无效却一致的估计量。同样地,这违背了通常 OLS 结果:如果误差是序列无关的,那么 y_{it} 对 $y_{i,t-1}$ 的回归得出一致估计值(尽管在小样本中出现偏倚)。

其次,考察组内估计量,即($y_{it}-\bar{y}_i$)对($y_{i,t-1}-\bar{y}_{i,-1}$)与($\mathbf{x}_{it}-\bar{\mathbf{x}}_i$)进行回归。该回归具有误差项($\varepsilon_{it}-\bar{\varepsilon}_i$)。现在,由式(22.32)知,$y_{it}$ 与 ε_{it} 是相关的,所以 $y_{i,t-1}$ 与 $\varepsilon_{i,t-1}$ 是相关的,从而与 $\bar{\varepsilon}_i$ 相关。然而,这蕴含回归元($y_{it}-\bar{y}_i$)[1]与误差($\varepsilon_{it}-\bar{\varepsilon}_i$)是相关的。因此,组内模型的 OLS 估计会产生非一致参数估计值,因为回归元与误差项是相关的。一致性要求,$\bar{\varepsilon}_i$ 相对 ε_{it} 而言变得非常小,这要求 $T\to\infty$,在长面板数据情况下,会出现此条件,但在短面板数据下则不会。重要的参考文献是尼克尔(Nickell, 1981)。

由第 21 章给出的随机效应估计量也会产生非一致性,因为这是组内估计与组间估计量的线性组合。对于随机效应模型,当 $\varepsilon_{it}\sim\mathcal{N}[0,\sigma^2]$ 时,安德森和萧政(Anderson and Hsiao, 1981)反而考察 ML 估计;也可参见巴尔加瓦和萨根(Bhargava and Sargan, 1983)。在短面板中,MLE 的分布依赖于对 y_{i1}[2]做出的假设,即因变量的初始值。安德森和萧政(Anderson and Hsiao, 1981)对下述初始条件假设进行了辨析:(1) 固定初始观测值;(2) 具有共同均值的随机初始观测值;

[1] 原著中这里为"$y_{i,t-1}-\bar{y}_i$",依据上下文判断,应为"$y_{it}-\bar{y}_i$"。——译者注

[2] 原著中这里为"y_{i0}",依据上下文判断,应为"y_{i1}"。——译者注

(3) 具有不同均值的随机初始观测值;(4) 具有平稳分布的随机初始观测值。

一阶差分估计量也是非一致的,但是 IV 变形产生了一致估计值。现在,我们阐述该估计量。

22.5.3 阿雷拉诺—邦德估计量

模型(22.32)导致了一阶差分模型[1]:

$$y_{it}-y_{i,t-1}=\gamma(y_{i,t-1}-y_{i,t-2})+(\mathbf{x}_{it}-\mathbf{x}_{i,t-1})'\boldsymbol{\beta}+(\varepsilon_{it}-\varepsilon_{i,t-1}),\quad t=3,\cdots,T \tag{22.34}$$

由式(22.32)知,因为 $y_{i,t-1}$ 与 $\varepsilon_{i,t-1}$ 是相关的,因而式(22.34)中的回归元($y_{i,t-1}-y_{i,t-2}$)与误差($\varepsilon_{it}-\varepsilon_{i,t-1}$)是相关的,所以该 OLS 估计量是非一致的。

安德森和萧政(Anderson and Hsiao, 1981)提出利用 $y_{i,t-2}$ 作为($y_{i,t-1}-y_{i,t-2}$)的工具估计式(22.34)的工具变量估计量。这是有效工具,因为一旦假定误差 ε_{it} 是序列无关的,$y_{i,t-2}$ 与($\varepsilon_{it}-\varepsilon_{i,t-1}$)就是无关的。进一步地,由于 $y_{i,t-2}$ 与($y_{i,t-1}-y_{i,t-2}$)是相关的,所以它是一个好工具。此方法需要每个个体有 3 个时期数据可利用。一种可供选择的方式是,使用 $\Delta y_{i,t-2}$ 作为关于 $\Delta y_{i,t-1}$ 的工具。这将需要 4 个时期数据。安德森和萧政(Anderson and Hsiao, 1981)阐述的结果表明,利用作为 $\Delta y_{i,t-2}$ 工具的 IV 估计量比利用 $y_{i,t-2}$ 作为 IV 工具的估计量更有效,如同通常情况 $\gamma>0$ 的工具一样。上述两种之一情况下,($\mathbf{x}_{it}-\mathbf{x}_{i,t-1}$)作为其自身工具。

利用因变量的额外滞后作为工具,获得更有效估计是可能的。例如,$y_{i,t-2}$ 和 $y_{i,t-3}$ 都可作为工具。于是,此模型是过度识别的,因此可通过 2SLS 或面板 GMM 加以估计。进一步地,可利用的工具个数越大,在时间 t 观测的因变量就越接近于最终时期 T。在第 3 个时期,仅有 y_{i1} 可作为工具,在第 4 个时期 y_{i1} 和 y_{i2} 都可作为工具,在第 5 个时期,则有 y_{i1}、y_{i2} 以及 y_{i3} 都可作为工具。霍尔茨·埃金等人(Holtz-Eakin et al., 1988)、阿雷拉诺和邦德(Arellano and Bond, 1991)都曾提出利用这些较广泛的非平衡工具集合的面板 GMM 估计量。

微观经济计量学文献将上述得到的面板 GMM 估计量称为阿雷拉诺—邦德估计量。一般方法已经由 22.4.2 节阐述,那里并没有以显性方式介绍其动态特性。此估计量是:

$$\hat{\boldsymbol{\beta}}_{\mathrm{AB}}=\left[\left(\sum_{i=1}^{N}\widetilde{\mathbf{X}}_i'\mathbf{Z}_i\right)\mathbf{W}_N\left(\sum_{i=1}^{N}\mathbf{Z}_i'\widetilde{\mathbf{X}}_i\right)\right]^{-1}\left(\sum_{i=1}^{N}\widetilde{\mathbf{X}}_i'\mathbf{Z}_i\right)\mathbf{W}_N\left(\sum_{i=1}^{N}\mathbf{Z}_i'\tilde{\mathbf{y}}_i\right) \tag{22.35}$$

其中,$\widetilde{\mathbf{X}}_i$ 表示$(T-2)\times(K+1)$阶矩阵,其第 t 行为($\Delta y_{i,t-1},\Delta\mathbf{x}_{it}'$), $t=3,\cdots,T$, $\tilde{\mathbf{y}}_i$ 表示$(T-2)\times 1$ 维向量,第 t 行为 Δy_{it},而 $\mathbf{Z}_i$ 表示$(T-2)\times r$ 阶工具矩阵:

$$\mathbf{Z}_i=\begin{bmatrix}\mathbf{z}_{i3}' & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \mathbf{z}_{i4}' & & \vdots\\ \vdots & & \ddots & \mathbf{0}\\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{z}_{iT}'\end{bmatrix} \tag{22.36}$$

[1] 原文中该式条件为"$t=2,\cdots,T$",依据上下文判断,应为"$t=3,\cdots,T$"。——译者注

这里,$\mathbf{z}_{it}'=[y_{i,t-2},y_{i,t-3},\cdots,y_{i1},\Delta\mathbf{x}_{it}']$。此外,$\mathbf{x}_{it}$或$\Delta\mathbf{x}_{it}$的滞后项能用作工具,而且对于适度的或大$T$来说,可能存在$y_{it}$的最大滞后作为工具,例如并不大于$y_{i,t-4}$。两阶段LS与两步GMM对应于不同的加权矩阵(参见22.2.3节)。

此方法很容易适应AR(p)模型,只是式(22.32)中的$\gamma y_{i,t-1}$代替$\gamma_2 y_{i,t-1}+\gamma_2 y_{i,t-2}+\cdots+\gamma_p y_{i,t-p}$,尽管为了使一致估计可行,就需要多于3个时期数据。

22.3节的实证例子本质上是阿雷拉诺—邦德估计例子,因为一阶差分模型可通过含有滞后回归元用作工具的IV加以估计。

阿和施密特(Ahn and Schmidt, 1995)注意到,利用额外矩条件可能获得更有效估计。考察式(22.32)的纯时间序列形式,其中$\boldsymbol{\beta}=\mathbf{0}$,同时做出标准假设,即$\varepsilon_{it}$与$\alpha_i$、$\varepsilon_{is}$以及初始观测值$y_{i1}$是无关的,对于$s\neq t$。阿雷拉诺—邦德估计量使用了矩条件$\mathrm{E}[y_{is}\Delta u_{it}]=0$,对于$s\leqslant t-2$,其中,$u_{it}=\varepsilon_{it}+\alpha_i$。阿和施密特(Ahn and Schmidt, 1995)通过利用额外矩条件$\mathrm{E}[u_{iT}\Delta u_{it}]=0$得到了更有效估计量。他们已经证明,这种估计量有效运用了二阶矩条件,它渐近等价于张伯伦(Chamberlain, 1982,1984)的最优最小距离估计量。

额外假设导致了额外矩条件,从而产生更有效估计。一旦假定ε_{it}同方差性,如果$\mathrm{V}[\varepsilon_{it}]=\mathrm{V}[\varepsilon_{is}]$,那么$\mathrm{E}[\bar{u}_i\Delta u_{it}]=0$[参见阿和施密特(Ahn and Schmidt, 1995)]。阿雷拉诺和博韦(Arellano and Bover, 1995)提出了对于$s\leqslant t-1$,利用条件$\mathrm{E}[u_{it}\Delta y_{is}]=0$。布伦德尔和邦德(Blundell and Bond, 1998)考察了这些假设和额外假设,同时证明其益处是很大的,特别是当γ是高的且T是小的时候。阿雷拉诺和霍诺尔(Arellano and Honore, 2001)阐述了可能做出的各种假设以及相应的可用于估计的矩条件。

萧政、佩萨兰和塔赫米斯吉奥卢(Hsiao, Pesarn, and Tahmiscioglu, 2002)提出变换ML估计量(**transformed ML estimator**)。假定ε_{it}服从iid $\mathcal{N}[0,\sigma^2]$,可以放松这一假设。其原因不是源自基于$\varepsilon_{i1},\cdots,\varepsilon_{iT}$似然,而是源自基于误差差分$\Delta\varepsilon_{i1},\cdots,\Delta\varepsilon_{iT}$似然。对于纯时间序列AR(1)模型,$\Delta\varepsilon_{it}=\Delta y_{it}-\gamma\Delta y_{i,t-1}$,对于$t>1$。$\Delta\varepsilon_{i1}$的密度依赖于对初始条件所做出的假设:或者$\Delta\varepsilon_{i1}=\Delta y_{i1}$或者$\Delta\varepsilon_{i1}=\Delta y_{i1}-b$,其中,$b=\mathrm{E}[\Delta y_{i1}]$表示待估的另外参数。即使$\varepsilon_{it}$是非正态的,所得到的估计量仍是保持一致性的拟MLE。如果ε_{it}服从iid$[0,\sigma^2]$,那么与前面的GMM估计量相比,变换MLE更为有效。

22.5.4 *协方差结构估计*

协方差结构是对回归误差的协方差矩阵的结构加以设定的模型。一些应用包括误差动态特性与测量误差的结构。目的是要估计结构参数。

举一个例子,假定y_{it}是由含有MA(1)误差的随机效应模型所生成的,因此:

$$y_{it}=\alpha_i+\varepsilon_{it}+\phi\varepsilon_{i,t-1}$$

其中,$\alpha_i\sim[0,\sigma_\alpha^2]$,$\varepsilon_{it}\sim[0,\sigma_\varepsilon^2]$,而且$|\phi|<1$。从而,其自协方差$\gamma_j=\mathrm{Cov}[y_{it},y_{it-j}]$,满足$\gamma_0=\sigma_\alpha^2+(1+\phi^2)\sigma_\varepsilon^2$,$\gamma_1=\sigma_\alpha^2+\phi\sigma_\varepsilon^2$,并且$\gamma_j=\sigma_\alpha^2$,对于$j\geqslant 2$。当$T=3$时,这些式子就产生了给定自协方差估计值$\hat{\gamma}_0$、$\hat{\gamma}_1$、$\hat{\gamma}_2$时的估计值$\hat{\sigma}_\alpha^2$、$\hat{\sigma}_\varepsilon^2$以及$\hat{\phi}$。当$T>3$

时，那些模型是过度识别的，因为要估计 3 个以上协方差，却仅有 3 个方差参数。一个明显的估计量是最小距离估计量。

通常，设 $\boldsymbol{\theta}$ 表示 q 个结构参数，同时假定 $\mathbf{g}(\boldsymbol{\theta})=\boldsymbol{\gamma}$，其中，$\boldsymbol{\gamma}=[\gamma_0,\cdots,\gamma_{T-1}]'$ 表示 $T\geqslant q$ 个自协方差向量。于是，最小距离估计量是对：

$$Q_N(\boldsymbol{\theta})=(\hat{\boldsymbol{\gamma}}-\mathbf{g}(\boldsymbol{\theta}))'\mathbf{W}_N(\hat{\boldsymbol{\gamma}}-\mathbf{g}(\boldsymbol{\theta})) \tag{22.37}$$

求极小值，其中，$\hat{\boldsymbol{\gamma}}=[\hat{\gamma}_1,\cdots,\hat{\gamma}_{T-1}]'$，且：

$$\hat{\gamma}_j=[N(T-j)]^{-1}\sum_{t=j+1}^{T}\sum_{i=1}^{N}(y_{it}-\bar{y}_i)(y_{i,t-j}-\bar{y}_{t-j}) \tag{22.38}$$

而 $\bar{y}_{t-j}=N^{-1}\sum_i y_{i,t-j}$，6.7 节已经提供了加权矩阵 $\mathbf{W}_N$ 以及有关 MD 估计的进一步详细内容。此模型的约束，可通过由 6.7 节给出的卡方检验统计量来加以检验。就此范围来说，这种讨论已经对协方差平稳性施加了约束。更一般地，人们能够允许 $\gamma_{tj}\neq\gamma_{sj}$，对于 $t\neq s$，其中，$\gamma_{tj}=\mathrm{Cov}[y_{it},y_{i,t-j}]$。于是，$\boldsymbol{\gamma}$ 拥有 $T(T+1)/2$ 个元素 γ_{tj}，$t=j+1,\cdots,T$，$j=0,\cdots,T-1$。该平稳性假设本身是可检验的假设。此外，一些回归元可通过用残差 $y_{it}-\mathbf{x}_{it}'\boldsymbol{\beta}$ 代替 y_{it} 而得以并入。

阿博特和卡德（Abowd and Card，1989）提供了早期的这一方法应用于工资与工时的联合建模。奥尔顿吉和西格尔（Altonji and Segal，1996）证明了，在有限样本中，最优 MD 估计量是相当有偏的（参见 6.3.5 节）。许多应用都是利用工资进行建模；参见贝克和索伦（Baker and Solon，2003）最新例子。

MD 方法更适合于对协方差结构进行估计。面板数据集可能是很大的，但首先借助于对协方差加以估计，该估计简化成对式(22.37)求极小值。其他一些方法是可行的。特别地，参见麦柯迪（McCurdy，1982b），他阐述了面板数据的博克斯—詹金斯形式模型。

22.5.5 非平稳面板

有关单位根与非平稳的面板文献强调 N 和 T 都很大的面板。关于单位根检验，早期的重要论文是由莱文和林（Levin and Lin，1992）完成，但最终却由莱文、林和朱（Levin，Lin and Chu，2002）发表；佩萨兰和史密斯（Pesaran and Smith，1995）撰写了早期考察协整（**cointegration**）的论文。菲利普斯和穆恩（Phillips and Moon，1999）以及佩德罗尼（Pedroni，2004）都提供用于非平稳面板数据的一般推断理论。利用时序极限理论的分析是最简单的，其中比如说，首先固定 N，且 $T\to\infty$，随后 $N\to\infty$。更稳健的方法是，使用联合极限，其中，$T\to\infty$ 且 $N\to\infty$。最新文献评述，包括菲利普斯和穆恩（Phillips and Moon，2000）以及巴尔塔吉（Baltagi，2001，第 12 章）。

对短面板的非平稳数据来说，所做的研究还不多，有待于进一步探索研究。哈里斯和察韦里斯（Harris and Tzavalis，1999）考察了短面板的莱文和林（Levin and Lin，1992）单位根检验。设 $\hat{\gamma}$ 表示 AR(1) 固定效应模型 $y_{it}=\alpha_i+\gamma y_{i,t-1}+\varepsilon_{it}$ 中的 γ 组内估计值，其中，$\varepsilon_{it}\sim\mathrm{iid}\ \mathcal{N}[0,\sigma^2]$。我们考察单位根的零假设，因此 $\gamma=1$，并且没有截距 $\alpha_i=0$，这对应于哈密尔顿（Hamilton，1994，第 490 页）的纯时间序列的

第二种情况。在零假设下,当 T 固定且 $N\to\infty$时,单位根检验统计量是:

$$\frac{\sqrt{N}(\hat{\gamma}-1+3/(T+1))}{[3(17T^2-20T+17)]/[5(T-1)(T+1)^3]}\xrightarrow{d}\mathcal{N}[0,1]$$

这个统计量若有很大负值,则拒绝单位根假设。莱文和林(Levin and Lin, 1992)提供了另一些检验,诸如具有个体时间趋势的模型。

宾德、萧政和佩萨兰(Binder, Hsiao, and Pesaran, 2003)考察含有单位根与协整的固定效应动态模型的短面板估计。对单位根而言,阿雷拉诺—邦德估计量是非一致的,尽管阿和施密特(Ahn and Schmidt, 1995)已做出一些推广,但由22.5.3节结尾讨论的其他估计量却产生了一致估计量。宾德等人(Binder et al., 2003)曾提出拟ML估计量,当对单位根加以讨论时,该估计量在有限样本中表现良好。

22.6 差异中差分估计量

第25章将要阐述的评估文献关注于测算处理效应,在最简单情况下,如果处理发生,那么单个二值回归元的影响或边际效应等于1;如果处理不发生,那么回归元的影响或边际效应等于0。例如,关注内容在于测算政策变化(二值处理)对工资的效应,政策变化涉及变动税率或福利,或者某些人接收培训,而另一些人则没有。

在本节,我们涉及第25章与面板方法有关的方法。特别是,如果在处理前后有面板数据可以利用,同时并不是所有的个体者都接收处理,那么处理效应就能利用标准面板数据方法加以测算。于是,固定效应模型的一阶差分估计量就简化成简单的估计量,称为差异中差分估计量,这已在3.4.2节引入,并且将在25.5节继续研究。后一种估计量具有下述优点,即当存在重复横截面数据而不是面板数据可以利用时,同样可以运用它。然而,它确实依赖于经常不是以显性方式做出的模型假设。这里的研究遵循布伦德尔和麦柯迪(Blundell and MaCurdy, 2000)的线索展开。

22.6.1 *含有二值处理的固定效应*

设关注的二值回归元是:

$$D_{it}=\begin{cases}1, & \text{若个体 } i \text{ 在 } t \text{ 时期接受处理}\\0, & \text{其他}\end{cases}\tag{22.39}$$

假定关于 y_{it} 的固定效应模型满足:

$$y_{it}=\phi D_{it}+\delta_t+\alpha_i+\varepsilon_{it}\tag{22.40}$$

其中,δ_t 表示特定时间固定效应,而 α_i 表示特定个体固定效应。正如21.2.1节提及的,这等价于 y_{it} 对 D_{it} 与含有特定个体固定效应复杂情况的所有时间虚拟变量集合的回归。为了简单起见,没有其他回归元。

个体效应 α_i 可通过一阶差分加以剔除。于是，有：

$$\Delta y_{it}=\phi\Delta D_{it}+(\delta_t-\delta_{t-1})+\Delta\varepsilon_{it} \tag{22.41}$$

处理效应 ϕ 能借助于 Δy_{it} 对 ΔD_{it} 与所有时间虚拟变量集合的混合回归得到一致估计。

22.6.2 *差异中差分*

现在，考察只有两个时期的特定化情况。进一步地，假定处理仅仅发生在第 2 个时期，所以在第 1 个时期，对于所有个体，$D_{i1}=0$，并在第 2 个时期，对于已处理个体，$D_{i1}=1$，而对于未处理个体，$D_{i2}=0$。于是，下标 t 可从式(22.41)中省略，从而：

$$\Delta y_i=\phi D_i+\delta+v_i \tag{22.42}$$

其中，D_i 表示二值处理变量，表明个体是否接收处理。

处理效应可通过 Δy 对截距与二值变量回归元 D 进行回归而得以估计。将 $\Delta\bar{y}^{\text{tr}}$定义成已处理($D_i=1$)的样本平均，而将 $\Delta\bar{y}^{\text{nt}}$定义成未处理($D_i=0$)的样本平均。于是，估计量简化成：

$$\hat{\phi}=\Delta\bar{y}^{\text{tr}}-\Delta\bar{y}^{\text{nt}} \tag{22.43}$$

此估计量称为差异中差分估计量(**differences-in-differences estimator**，记为 **DID**)，因为人们对已处理组与未处理组的时间差异进行估计，然后对时间差异取差分。

此估计量由于直观简单而引人注目。另外，它能从面板数据推广到两个时期各个横截面数据均可利用的情况。在第 2 个时期，计算已处理组与未处理组的平均值 $\bar{y}_2^{\text{tr}}$ 与 $\bar{y}_2^{\text{nt}}$。类似地，可计算出第 1 个处理前时期的平均值 $\bar{y}_1^{\text{tr}}$ 与 $\bar{y}_1^{\text{nt}}$。这里假定第 1 个时期中个体是否适宜处理是可识别的。例如，若处理仅仅应用于妇女，而且可以利用性别数据，则很容易实施。于是，计算：

$$\hat{\phi}=(\bar{y}_2^{\text{tr}}-\bar{y}_1^{\text{tr}})-(\bar{y}_2^{\text{nt}}-\bar{y}_1^{\text{nt}}) \tag{22.44}$$

举一个例子，倘若适宜处理组的年工资在处理前为 10 000，而在处理后为 13 000，则 $\bar{y}_2^{\text{tr}}-\bar{y}_1^{\text{tr}}=3\,000$。类似地，如果不适宜处理组的年工资在处理前为 15 000，而在处理后为 17 000，那么 $\bar{y}_2^{\text{nt}}-\bar{y}_1^{\text{nt}}=2\,000$。从而，处理效应的估计量 $\hat{\phi}$等于 $3\,000-2\,000=1\,000$。

22.6.3 *差异中差分的假设基础*

前面的 DID 估计量公式为了得到 ϕ 的一致估计，已做出了明显的基本假设。

首先，假定时间效应 δ_t 对于不同的已处理个体与未处理个体来说都是共同的。例如，时间趋势可因性别而不同，在此情况下，若处理依赖于性别，识别 ϕ 就会有问题。不论是使用面板数据，还是使用横截面数据，都需要共同趋势假设。

其次，若使用横截面数据，则假定已处理组与未处理组的合成部分在变动前后均是稳定的。就面板数据而言，进行差分会剔除固定效应。就重复横截面数据而言，最初模型(22.40)蕴含着 $\bar{y}_t^{\text{tr}}=\phi+\delta_t+\bar{\alpha}_t^{\text{tr}}+\bar{\varepsilon}_t^{\text{tr}}$ 与 $\bar{y}_t^{\text{nt}}=\delta_t+\bar{\alpha}_t^{\text{nt}}+\bar{\varepsilon}_t^{\text{nt}}$。倘若处理只

在第 2 个时期发生,由此可得:

$$\phi=(\bar{y}_2^{tr}-\bar{y}_1^{tr})-(\bar{y}_2^{nt}-\bar{y}_1^{nt})+(\bar{\alpha}_2^{tr}-\bar{\alpha}_1^{tr})-(\bar{\alpha}_2^{nt}-\bar{\alpha}_1^{nt})+v$$

其中,$v=(\bar{\varepsilon}_2^{tr}-\bar{\varepsilon}_1^{tr})-(\bar{\varepsilon}_2^{nt}-\bar{\varepsilon}_1^{nt})$[1]。如果 $\text{plim}(\bar{\alpha}_2^{tr}-\bar{\alpha}_1^{tr})=0$ 且 $\text{plim}(\bar{\alpha}_2^{nt}-\bar{\alpha}_1^{nt})=0$,那么出现式(22.44)中 $\hat{\phi}$ 的一致性。若处理指派是随机的,正是此种情况。可是,事实并不经常如此。

22.6.4 *更多模型*

在实际应用中,人们会使用更多模型。一个明显推广是,包括一些回归元而不只是处理指示变量与时间虚拟变量。通过对数据进行分组,特定个体效应至少允许在不同组平均值上各不相同,一般方法是估计:

$$y_{igt}=\phi D_{igt}+\delta_t+\alpha_i+\varepsilon_{it}$$

其中,g 表示第 g 个组。

在 DID 估计的经典例子中,卡德(Card, 1990)研究了从古巴突然涌入迈阿密的移民对低收入工人失业的效应。这个例子同样被安格里斯特和克鲁格(Angrist and Krueger, 1999)评述。阿西和英伯斯(Athey and Imbens, 2002)讨论了非线性模型的推广。

22.7 重复横截面与伪面板

面板数据的重要潜在优势,源自不同时间能观测到对象目标。这使得控制不可观测个体异质性、初始条件差异以及结果的动态相依性成为可能。然而,在许多情况下,并不可以利用名副其实的面板数据。

22.7.1 *重复横截面*

我们考察如下问题:当数据是几个重复横截面时,这里的重复横截面来自一系列独立样本调查的响应,独立性意味着每一个对象目标只出现在一个调查之中。一个例子是英国家庭支出调查,它收集了大量家庭支出数据年度样本,但每一年都调查不同家庭。并且,如果仅有非常短面板是可以利用的(比如 $T=2$),那么来自重复横截面的数据就引人注目,假如它们可生成较大且较丰富的样本。

对于随机效应模型来说,重复横截面数据并没有提出什么挑战。人们可直接实施 y_{it} 对 $\mathbf{x}_{it}$ 混合回归(参见 21.5 节),而统计推断实际上得到简化,由于此处误差既关于 i 又关于 t 均是独立的,所以只有针对异方差性,才需要加以修正。

然而,对于固定效应来说,混合回归会导致非一致参数估计值。进一步地,如果个体仅仅在一个时点上是可观测的,那么一些可供选择方法,诸如组内或一阶差分估计均是不可行的。在本节,重复横截面数据用于建立伪面板(**pseudo panel**)或

[1] 原著中该式为 $v=(\bar{\varepsilon}_2^{nt}-\bar{\varepsilon}_1^{nt})-(\bar{\varepsilon}_2^{nt}-\bar{\varepsilon}_1^{nt})$,依据上下文判断,应为 $v=(\bar{\varepsilon}_2^{tr}-\bar{\varepsilon}_1^{tr})-(\bar{\varepsilon}_2^{nt}-\bar{\varepsilon}_1^{nt})$。——译者注

综合面板数据(**synthetic panel data**),这类数据拥有真正面板数据的某些优点,最值得注意的是控制固定效应的能力。特殊情况是 22.6 节已阐述的 DID 估计量。

22.7.2 *伪面板*

布朗宁、迪顿和艾里什(Browning, Deaton, and Irish, 1985)在基于英国家庭支出调查的实证研究时,考察了用于分析重复横截面数据的一些方法。他们提出将个体水平数据转换成组类水平(**cohort-level**)数据。尽管个体家庭支出不能随时间流逝而加以叠放,但对由一些个体构成的组类却可能这样做。

组类(**cohort**)被定义成“具有固定从属关系的组,那些在调查中可被排列起来检查且可识别的个体”[迪顿(Deaton,1985,第 109 页)]。一个例子是年龄组类,诸如在 1965～1970 年之间出生的男性。对于大样本,连续不断的调查将会生成每个组类成员的随机样本。

组类样本平均时间序列能够形成回归模型的基础。基于组类数据的综合面板能否代替真正面板数据是一个关键问题。重复横截面专题研究此类模型的推断方法。这里,我们关注静态伪面板模型。科拉多(Collado, 1997)与吉尔马(Girma, 2000)还考察了动态情况。

起点是含有个体固定效应 α_i 的静态线性回归,它建立在 T 个连续不断的横截面基础上:

$$y_{it}=\alpha_i+\mathbf{x}_{it}'\boldsymbol{\beta}+u_{it}, \quad t=1,\cdots,T \tag{22.45}$$

假定解释变量关于关注参数 $\boldsymbol{\beta}$ 是强外生的,因而 $\mathrm{E}[\mathbf{x}_{it}'u_{is}]=\mathbf{0}, \forall t,s$。为了简单起见,假定对每个横截面都有 N 个观测值可以利用。每一个个体仅仅在一个时期可以观测到,所以特定个体效应 α_i 不能借助于对个体水平数据进行差分而剔除。

设 g 表示下述随机变量,对于每个 i,它决定组类从属关系,使得 i 属于 c 类,当且仅当 g_i 属于集合 I_c。假定存在 C 个组类,而 c 表示组类下标,$c=1,\cdots,C$。若取以 g_i 为条件的期望,得出:

$$\mathrm{E}[y_{it}\mid g_i\in I_c]=\mathrm{E}[\alpha_i\mid g_i\in I_c]+\mathrm{E}[\mathbf{x}_{it}'\mid g_i\in I_c]\boldsymbol{\beta}+\mathrm{E}[u_{it}\mid g_i\in I_c] \tag{22.46}$$

这就生成了模型(22.45)的组类总体(**cohort population**)形式,它由

$$y_{ct}^*=\alpha_c^*+\mathbf{x}_{ct}^{*\prime}\boldsymbol{\beta}+u_{ct}^* \tag{22.47}$$

给出。其中,“ * ”号表示不能观测到的组类总体平均。例如,$y_{ct}^*=\mathrm{E}[y_{it}\mid g_i\in I_c]$。

参数 $\alpha_c^*=\mathrm{E}[\alpha_i\mid g_i\in I_c]$=表示组类固定效应(**cohort fixed effect**)。在固定效应情况下做出的一个重要假设是,总体是平稳的,因此,α_c^* 能被假定成随时间变化而为常值。这在性质上类似于 22.6.3 节结尾做出的 DID 估计量一致性需要的假设。在通常的弱外生性假设下,$\mathrm{E}[u_{ct}^*\mid \mathbf{x}_{ct}^*]=0$。不过,如果最初模型(22.45)中的 α_i 与 $\mathbf{x}_{it}$ 是相关的,那么不可观测固定效应 α_c^* 将与 $\mathbf{x}_{ct}^*$ 相关。进行估计时,就需要控制固定效应。

在实际应用中,组类总体均值是不能观测到的,不过,我们以组类时间平均 $\bar{y}_{ct}$ 与 $\bar{\mathbf{x}}_c$ 来开始研究。于是,此回归为:

$$\bar{y}_{ct}=\bar{\alpha}_c+\bar{\mathbf{x}}_{ct}'\boldsymbol{\beta}+\bar{u}_{ct},\quad c=1,\cdots,C,\ t=1,\cdots,T \tag{22.48}$$

上述步骤引入了额外误差来源,因为 $\bar{y}_{ct}$ 与 $\bar{\mathbf{x}}_{ct}$ 都是组类总体平均的误差污染估计值,即:

$$\begin{aligned}\bar{y}_{ct}&=y_{ct}^*+\xi_{ct}\\ \bar{\mathbf{x}}_{ct}&=\mathbf{x}_{ct}^*+\boldsymbol{v}_{ct}\end{aligned} \tag{22.49}$$

如果测量误差是非常小的,这归因于每个时期每一个组类的观测值数目都相当大,那么 $\bar{y}_{ct}\simeq y_{ct}^*$ 且 $\bar{\mathbf{x}}_{ct}=\mathbf{x}_{ct}^*$,从而忽略其测量误差。$\boldsymbol{\beta}$ 的一致估计值能借助于式(22.48)的组内估计来获得,也就是说,$(\bar{y}_{ct}-\bar{y}_c)$ 对 $(\bar{\mathbf{x}}_{ct}-\bar{\mathbf{x}}_c)$ 进行 OLS 回归,其中,$\bar{y}_c=T^{-1}\sum_t\bar{y}_{ct}$,$\bar{\mathbf{x}}_c=T^{-1}\sum_t\bar{\mathbf{x}}_{ct}$。

不幸的是,测量误差往往由于太大而不能被忽略。于是,当 $\bar{\alpha}_c$ 为随机效应时,式(22.48)的组内估计或式(22.48)的 OLS 估计均会产生 $\boldsymbol{\beta}$ 的非一致估计值。相反,需要使用变量误差估计量。由于个体水平数据会产生测量误差矩的必要估计值,所以此处能实施这类估计量,参见 26.3.3 节。

22.7.3 伪面板的测量误差估计量

对测量误差的经典求解是使用重复观测值来估计测量误差的协方差,然后在应用最小二乘法之前使用这些估计值去"校正"污染误差变量的样本矩(参见 26.4.4 节)。迪顿(Deaton, 1985)提出在当前背景下使用此方法。

假定个体观测值满足下述方程:

$$\begin{aligned}y_{it}&=y_{ct}^*+\varepsilon_{it}\\ \mathbf{x}_{it}&=\mathbf{x}_{ct}^*+\boldsymbol{\eta}_{it}\end{aligned}$$

背景设置类似于 26.2.1 节的背景,只是因变量还存在测量误差,同时假定对于给定 c 组类中的任何个体,有:

$$\begin{bmatrix}\varepsilon_{it}\\ \boldsymbol{\eta}_{it}\end{bmatrix}\sim\text{iid}\left[\begin{bmatrix}0\\ \mathbf{0}\end{bmatrix},\begin{bmatrix}\sigma_0^2 & \boldsymbol{\sigma}_{01}'\\ \boldsymbol{\sigma}_{01} & \boldsymbol{\Sigma}\end{bmatrix}\right]$$

$(\boldsymbol{\Sigma},\boldsymbol{\sigma}_{01})$ 的样本估计值记为 $(\hat{\boldsymbol{\Sigma}},\hat{\boldsymbol{\sigma}}_{01})$,给定 $(\bar{y}_{ct},\bar{\mathbf{x}}_{ct})$ 时可利用所有个体水平数据来获得。记 $\mathbf{d}_c$ 表示对应于固定效应 α_c^*(参见 21.2.1 节)虚拟变量的 $C\times1$ 维列向量,很明显这是一个不受估计误差限制的回归元向量。于是,倘若 T 充分大且其逆存在,当 $CT\to\infty$ 时,则回归:

$$\begin{bmatrix}\hat{\bar{\alpha}}_{ct}\\ \hat{\bar{\boldsymbol{\beta}}}_{ct}\end{bmatrix}=\left(\sum_{c=1}^{C}\sum_{t=1}^{T}\begin{bmatrix}\mathbf{d}_c'\mathbf{d}_c & \mathbf{d}_c'\bar{\mathbf{x}}_{ct}\\ \bar{\mathbf{x}}_{ct}'\mathbf{d}_c & \bar{\mathbf{x}}_{ct}'\bar{\mathbf{x}}_{ct}-\hat{\boldsymbol{\Sigma}}\end{bmatrix}\right)^{-1}\left[\sum_{c=1}^{C}\sum_{t=1}^{T}\begin{pmatrix}\mathbf{d}_c'\bar{y}_{ct}\\ \bar{\mathbf{x}}_{ct}'\mathbf{d}_c-\hat{\boldsymbol{\sigma}}_{01}\end{pmatrix}\right] \tag{22.50}$$

将给出组类回归的一致估计。这个估计量与 26.3.4 节所给出的一样,只需进行适当修改,因为此处 $\bar{y}_{ct}$ 也是含有误差且可以简化的测量,其原因在于仅有回归元的子集是含有误差的测量。维比克和尼吉曼(Verbeek and Nijman, 1992)对抽样性质提供了更详细的讨论,而迪顿(Deaton, 1985)阐述了方差估计。还可以参见维比克(Verbeek, 1995)。

上述估计量通过估计最小二乘法虚拟变量模型，本质上控制组类固定效应，一旦利用由 26.3.4 节给出的估计量，可借助复制数据调整测量误差。

科拉多(Collado，1997)考察了一种利用一阶差分剔除组类效应的可供选择的方法，然后通过工具变量估计对测量误差加以控制，而 26.3.2 节将给出一种可供选择的识别测量误差的策略。

把式(22.49)代入式(22.47)，得出：

$$\bar{y}_{ct}-\xi_{ct}=\alpha_c^*+(\bar{\mathbf{x}}_{ct}'-\boldsymbol{\upsilon}_{ct}')\boldsymbol{\beta}+u_{ct}^*$$
$$\bar{y}_{ct}=\alpha_c^*+\bar{\mathbf{x}}_{ct}'\boldsymbol{\beta}+w_{ct}$$

其中，误差 $w_{ct}=u_{ct}^*+\boldsymbol{\upsilon}_{ct}'\boldsymbol{\beta}+\xi_{ct}$。运用一阶差分剔除 α_c^*，得到：

$$\Delta\bar{y}_{ct}=\Delta\,\bar{\mathbf{x}}_{ct}'\boldsymbol{\beta}+\Delta w_{ct},\quad t=2,\cdots,T \tag{22.51}$$

现在，由于测量误差项的缘故，解释变量 $\Delta\,\bar{\mathbf{x}}_{ct}'$ 将与 Δw_{ct} 相关，从而若应用最小二乘法，将产生非一致估计。一致估计能通过基于外生变量滞后项即 $\bar{\mathbf{x}}_{c,t-1}'$ 的 IV 估计来获得。这种方法具有下述优点：它可以推广到含有滞后因变量的模型上。详细内容参见科拉多(Collado，1997)。

22.8 混合线性模型

被经济计量学家称为随机效应的模型，只是将截距系数设定成随机的。更丰富的随机效应模型，广泛用于应用统计学的其他领域，这类模型额外地允许斜率参数是随机的。在本节，我们阐述混合线性模型——也称为混合效应模型、分层模型或多水平线性模型(参见第 24 章)、随机效应模型以及方差成分模型。

这些模型应用在使混合 OLS 估计量仍为一致的背景下。特别地，不存在固定效应。由于混合线性模型框架提供了足够多的结构，以致允许借助于可行 GLS 进行估计，其估计值更为有效。

22.8.1 *混合线性模型*

混合线性模型(**mixed linear model**)设定：

$$y_{it}=\mathbf{z}_{it}'\boldsymbol{\beta}+\mathbf{w}_{it}'\boldsymbol{\alpha}_i+\varepsilon_{it} \tag{22.52}$$

其中，回归元 $\mathbf{z}_{it}$ 包括截距，$\mathbf{w}_{it}$ 表示可观测特征向量，$\boldsymbol{\alpha}_i$ 表示均值为 0 的随机向量，而 ε_{it} 表示误差项。此模型称为混合模型，因为它既有固定参数 $\boldsymbol{\beta}$，又有均值的随机参数或随机效应 $\boldsymbol{\alpha}_i$。

随机截距模型是 $y_{it}=\mathbf{z}_{it}'\boldsymbol{\beta}+\alpha_i+\varepsilon_{it}$，它是满足 $\mathbf{w}_{it}'\boldsymbol{\alpha}_i=\alpha_i$ 的式(22.52)的特殊情况。

式(22.25)的另一种特殊情况是随机系数模型或随机参数模型。在回归模型背景下，我们假定：

$$y_{it}=\mathbf{z}_{it}'\boldsymbol{\beta}_i+\varepsilon_{it}$$

是正规线性回归，只是回归参数向量现在依据：

$$\boldsymbol{\beta}_i=\boldsymbol{\beta}+\boldsymbol{\alpha}_i$$

随不同个体而变化，其中，$\boldsymbol{\alpha}_i$表示零均值随机向量。将它代入上式，得出 $y_{it}=\mathbf{z}_{it}'\boldsymbol{\beta}+\mathbf{z}_{it}'\boldsymbol{\alpha}_i+\varepsilon_{it}$，这是满足 $\mathbf{w}_{it}=\mathbf{z}_{it}$ 的式(22.52)。

许多应用处于随机截距模型与随机系数之间，其中，$\mathbf{w}_{it}$ 往往为 $\mathbf{z}_{it}$ 的子集。尤其是，标准混合 ANOVA 模型与随机 ANOVA 模型也是其特殊情况，其中，向量 $\mathbf{w}_{it}$ 的第 k 个分量为 0 或 1，这要依据各种可能的聚集数据模型而定。例如，$\mathbf{z}_{it}$ 的一个分量可能是种族或性别指示变量。于是，y_{it} 的条件均值会随着性别或种族而变化。还可以认为，y_{it} 的条件方差也随着性别或种族而变化，这能借助于包括 $\mathbf{w}_{it}$ 而获得。混合模型是 ANOVA 模型的分支。分层线性模型或多水平线性模型(参见 24.6.2 节)也能表述成式(22.52)的特殊情况。

22.8.2 *估计*

目标是估计固定回归参数 $\boldsymbol{\beta}$、关于 $\boldsymbol{\alpha}_i$ 与 ε_{it} 的分布方差以及协方差参数。此模型的早期研究之一是由林德利和史密斯(Lindely and Smith, 1972)给出的贝叶斯内容。他们的一般性研究的一个简单例子是含有 $y_{it}\sim\mathcal{N}[\mathbf{z}_{it}'\boldsymbol{\beta}_i,\sigma^2]$ 的随机系数模型，其中，$\boldsymbol{\beta}_i\sim\mathcal{N}[\boldsymbol{\gamma},\boldsymbol{\Gamma}]$。例如，有关线性面板数据模型的贝叶斯分析，参见库普(Koop, 2003)。

这里我们遵循经典方法(**classical approach**)，它是基于哈维尔(Harville, 1977)研究工作而展开的，哈维尔曾给出早期的参考文献。混合模型(22.52)可以被划分成确定性成分 $\mathbf{x}_{it}'\boldsymbol{\beta}$ 与随机成分 $\mathbf{w}_{it}'\boldsymbol{\alpha}_i+\varepsilon_{it}$。其随机假设包括回归元 $\mathbf{x}_{it}$ 与零均值随机成分 $\boldsymbol{\alpha}_i$ 及 ε_{it} 是独立的假设。因而，y_{it} 对 $\mathbf{x}_{it}$ 的混合 OLS 回归提供了一致估计值。我们基本上处于 21.5 节的领域之中，当对误差项 $\mathbf{w}_{it}'\boldsymbol{\alpha}_i+\varepsilon_{it}$ 的方差矩阵施加结构时，就具有可行的 GLS 估计。在本节，我们沿着两种不同的方法阐述可行 GLS 估计量，以便估计 $\boldsymbol{\alpha}_i$ 与 ε_{it} 的方差及协方差，并考察随机成分 $\boldsymbol{\alpha}_i$ 的预测。

若以通常方式对给定个体时不同时间的观测值加以组合，则式(21.52)变成：

$$\mathbf{y}_i=\mathbf{Z}_i\boldsymbol{\beta}+(\mathbf{W}_i\boldsymbol{\alpha}_i+\boldsymbol{\varepsilon}_i) \tag{22.53}$$

通常假设是，$\boldsymbol{\alpha}_i$ 与 $\boldsymbol{\varepsilon}_i$ 对于不同 i 是独立的，且它们相互之间是独立的，满足 $\boldsymbol{\alpha}_i\sim[\mathbf{0},\boldsymbol{\Sigma}_\alpha]$ 且 $\boldsymbol{\varepsilon}_i\sim[\mathbf{0},\boldsymbol{\Sigma}_\varepsilon]$，因此误差项满足：

$$\mathbf{W}_i\boldsymbol{\alpha}_i+\boldsymbol{\varepsilon}_i\sim[\mathbf{0},\boldsymbol{\Omega}_i=\mathbf{W}_i\boldsymbol{\Sigma}_\alpha\mathbf{W}_i'+\boldsymbol{\Sigma}_\varepsilon]$$

于是，可行 GLS 估计量是：

$$\hat{\boldsymbol{\beta}}_{\mathrm{FGLS}}=\Big[\sum_{i=1}^{N}\mathbf{Z}_i'\hat{\boldsymbol{\Omega}}_i^{-1}\mathbf{Z}_i\Big]^{-1}\sum_{i=1}^{N}\mathbf{Z}_i'\hat{\boldsymbol{\Omega}}_i^{-1}\mathbf{y}_i \tag{22.54}$$

其中，$\hat{\boldsymbol{\Omega}}_i$ 关于 $\boldsymbol{\Omega}_i$ 是一致的。

执行运算需要 $\boldsymbol{\Omega}_i$ 的一致估计值。较简单的随机截距情况已在 21.7 节中阐述过，在此情况下，存在几种不同方法一致估计 σ_α^2 与 σ_ε^2 的方差成分，只是具有一些复

杂情形诸如偏倚以及可能出现负的估计值。这里，估计 Σ_α 与 Σ_ε 时会引出类似问题。

我们在随机成分以外的正态分布假设基础上阐述两种估计量。例如，对于更一般模型：

$$\mathbf{y}=\mathbf{Z}\boldsymbol{\beta}+(\mathbf{W}\boldsymbol{\alpha}+\varepsilon) \tag{22.55}$$

其表述可借助于适当地叠放式(22.53)而获得。假定 $\boldsymbol{\alpha}\sim\mathcal{N}[\mathbf{0},\mathbf{G}]$，且 $\varepsilon\sim\mathcal{N}[\mathbf{0},\mathbf{R}]$，其中，在当前应用中，$\mathbf{G}$ 与 $\mathbf{R}$ 均是 Σ_α 与 Σ_ε 的函数。关于混合模型的可行估计量是：

$$\hat{\boldsymbol{\beta}}_{\mathrm{FGLS}}=[\mathbf{Z}'\hat{\mathbf{V}}^{-1}\mathbf{Z}]^{-1}\mathbf{Z}'\hat{\mathbf{V}}^{-1}\mathbf{y}$$

其中，$\hat{\mathbf{V}}$ 关于 $\mathbf{V}=\mathrm{V}[\mathbf{W}\boldsymbol{\alpha}+\varepsilon]=\mathbf{WGW}'+\mathbf{R}$ 是一致的。参见斯瓦米(Swamy，1970)。

获得 $\hat{\mathbf{V}}$ 的一种明显方法是极大似然法。基于多变量正态的对数似然，即剔除 $\boldsymbol{\beta}$ 后等于 GLS 估计量，是：

$$\ln \mathrm{L}(\mathbf{G},\mathbf{R})=-\frac{1}{2}\ln|\mathbf{V}|-\frac{NT}{2}\ln \mathbf{r}'\mathbf{V}^{-1}\mathbf{r}-\frac{NT}{2}\left[1+\ln\left(\frac{2\pi}{NT}\right)\right]$$

其中，$\mathbf{r}=\mathbf{y}-\mathbf{Z}[\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}]^{-1}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{y}$，而 $|\mathbf{V}|$ 表示 $\mathbf{V}$ 的行列式。针对 $\mathbf{G}$ 与 $\mathbf{R}$ 中的参数求极大值，得出 $\hat{\mathbf{V}}=\mathbf{W}\hat{\mathbf{G}}\mathbf{W}'+\hat{\mathbf{R}}$。

方差成分的 ML 估计弱点是，它们在小样本中是有偏的。例如，对于含有同方差误差的横截面线性回归来说，MLE $\hat{\sigma}^2=N^{-1}\sum_i\hat{u}_i^2$ 是有偏的，不过，一种最好的方法是用$(N-K)$去除。对于模型(22.53)，自由度修正是由下述约束极大似然估计量提供的，该估计量极大化：

$$\ln \mathrm{L}_R(\mathbf{G},\mathbf{R})=-\frac{1}{2}\ln|\mathbf{V}|-\frac{NT-p}{2}\ln \mathbf{r}'\mathbf{V}^{-1}\mathbf{r}-\frac{NT-p}{2}\left[1+\ln\left(\frac{2\pi}{NT-p}\right)\right]-\frac{1}{2}\ln|\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}|$$

其中，p 表示 $\mathbf{Z}$ 的秩。有关 $\ln \mathrm{L}_R(\mathbf{G},\mathbf{R})$的动机，参见哈维尔(Harville，1977)。

举一个混合线性模型的实证例子，考察 21.3 节中的回归例子，该回归既允许截距是随机的，又允许斜率参数是随机的。于是，随机系数模型得出 lnhrs = 7.734 − 0.021lnwg，其斜率系数标准误差为 0.046(默认)，或者为 0.020(面板自助法)。此斜率系数与由表 21.2 给出的截然不同。

22.8.3 *预测*

除固定参数 $\boldsymbol{\beta}$ 与协方差参数之外，我们还希望预测随机参数 $\boldsymbol{\alpha}$。

给定 $\hat{\boldsymbol{\beta}}$ 与 $\hat{\boldsymbol{\alpha}}$ 的一致估计值时，关于 $\hat{\boldsymbol{\beta}}$ 与 $\hat{\boldsymbol{\alpha}}$ 的联合正规方程能写成：

$$\begin{bmatrix}\mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{Z} & \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{W}\\ \mathbf{W}'\hat{\mathbf{R}}^{-1}\mathbf{Z} & \mathbf{W}'\hat{\mathbf{R}}^{-1}\mathbf{W}+\hat{\mathbf{G}}^{-1}\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\boldsymbol{\alpha}}\end{bmatrix}=\begin{bmatrix}\mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{y}\\ \mathbf{W}'\hat{\mathbf{R}}^{-1}\mathbf{y}\end{bmatrix}$$

若求解 $\hat{\beta}$,则得出前面已给定的 $\hat{\beta}_{FGLS}$,而:

$$\hat{\alpha}=\hat{\mathbf{G}}\mathbf{W}\hat{\mathbf{V}}^{-1}(\mathbf{y}-\mathbf{Z}'\hat{\beta})$$

在对于不同 i 具有独立性的情况下,这会得出 $\hat{\alpha}_i=\hat{\boldsymbol{\Sigma}}_\alpha\mathbf{W}_i'\mathbf{V}_i^{-1}(\mathbf{y}_i-\mathbf{Z}_i'\hat{\beta})$。如果方差矩阵是已知的,这就是最佳线性无偏预测量。

22.9 应用研究

面板 2SLS 估计量实际上能利用横截面数据的恰好 2SLS 程序加以估计(参见 22.2.5 节),尽管所计算的标准误差要求是面板稳健的。关于最优 GMM 估计量,可以利用统计软件包矩阵命令或诸如编程语言来执行运算。

一些统计软件越来越多地采用面板命令,这些面板命令会自动执行本章所述的估计量,包括最著名的阿雷拉诺—邦德估计量。

22.10 文献注释

本章涵盖了最近几本教科书都曾研究的面板数据方面活跃的研究领域,尤其是巴尔塔吉(Baltagi, 1995, 2001),萧政(Hsiao, 1986, 2003),李明宰(M-J. Lee, 2002)以及阿雷拉诺(Arellano, 2003)的书。更高等的一些方法是由马加什和塞韦斯特(Matyas and Sevestre, 1995)与阿雷拉诺和奥诺雷(Arellano and Honore, 2001)提供。

22.2 张伯伦(Chamberlain, 1982,1984)强调使用外生性假设。他运用了最小距离估计。后来文献使用了 GMM 方法。李明宰(M-J. Lee, 2002)和阿雷拉诺(Arellano, 2003)特别强调 GMM 估计。也可参见阿和施密特(Ahn and Schmidt, 1999)的综述。

22.4 豪斯曼和泰勒(Hausman and Taylor, 1981)的模型是引人注目的。借助于对一些回归元与特定个体效应不相关的假设,使得对时常值回归元的系数进行识别成为可能。

22.5 与由巴莱斯特和纳络夫(Balesta and Nerlove, 1966)开始的文献相比,线性动态模型的范围非常有限。更完整的讨论是由巴尔塔吉(Baltagi, 2001,第 8 章)、萧政(Hsiao, 2003,第 4 章)以及阿雷拉诺(Arellano, 2003,第 5~8 章)给出。阿雷拉诺—邦德(Arellano-Bond, 1991)估计量尤其流行,因为它建议含有固定效应的动态模型。

22.6 差异中差分方法因为其简单而极为流行。尽管它运用重复横截面数据而不是面板数据,但面板数据解释有助于做出明显的基本假设。伯特兰等人(Betrand et al., 2004)证明了利用 22.2.3 节的方法在个体水平上对时间序列相关性加以校正的重要性。

22.8 混合线性模型在统计学文献中特别流行。这种混合线性模型在经济计量学文献中较少使用,其原因是对时常值特定个体固定效应不愿意施加结构约束。

习　题

22-1　考察 22.2.1 节的面板 GMM 估计量。

(a) 证明，给定式(22.3)后对二次函数 $Q_N(\boldsymbol{\beta})$ 求关于 $\boldsymbol{\beta}$ 的极小值，得到给定 $Q_N(\boldsymbol{\beta})$ 时的面板 GMM 估计量，其中，$Q_N(\boldsymbol{\beta})$ 是运用求和记号表示的。

(b) 证明，这个估计量等价于由式(22.4)定义的估计量。

(c) 为了简单起见，假定式(22.4)中的矩阵 $\mathbf{Z}$ 与 $\mathbf{X}$ 均是非随机的，同时 $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{u}$，其中，$\mathbf{u}$ 具有均值 0 且方差 Ω。求式(22.4)估计量的有限样本方差矩阵，并将其与式(22.5)的渐近结果进行比较。

(d) 简化 $r=K$ 情况下的面板估计量。

22-2　考察面板数据模型，$y_{it}=\alpha+\beta x_{it}+\gamma w_{it}+u_{it}$，$i=1,\cdots,N$，$t=1,\cdots,T$，其中，为了简单起见，不存在特定个体效应。假定纯量回归元 x_{it} 与 u_{is} 是相关的，对于所有 t 与 s。对于下述每一种表述来说，β 与 γ 的一致估计是否是可行的？如果可行，在 22.2 节讨论的基础上，给出所有合适的工具。假定有三个时期的数据可以利用，因而 $T=3$，同时注意到，变量不可以用作所有年份的工具，并且，在不同年份中可利用不同工具。

(a) 回归元 w_{it} 满足求和假设 $\mathrm{E}[\sum_t w_{it}u_{it}]=0$。

(b) 回归元 w_{it} 满足同时期外生性假设 $\mathrm{E}[w_{it}u_{it}]=0$，$t=1,\cdots,3$。

(c) 回归元 w_{it} 满足弱外生性假设 $\mathrm{E}[w_{is}u_{it}]=0$，$s\leqslant t$，$t=1,\cdots,3$。

(d) 回归元 w_{it} 满足强外生性假设 $\mathrm{E}[w_{it}u_{it}]=0$，$s,t=1,\cdots,3$。

22-3　重述第三个问题，存在三个时期数据，现在考察面板模型 $y_{it}=\alpha_i+\beta x_{it}+\gamma w_{it}+u_{it}$，其中，$\alpha_i$ 表示固定效应，而且考察建立在一阶差分模型 $y_{it}-y_{i,t-1}=\beta(x_{it}-x_{i,t-1})+\gamma(w_{it}-w_{i,t-1})+(u_{it}-u_{i,t-1})$ 基础上的 IV 估计。

22-4　考察由 22.6 节阐述的差异中差分(DID)估计量。假定时间趋势项 $(\delta_t-\delta_{t-1})$ 对于已处理组与未处理组是不同的。

(a) 基于重复横截面数据的 ϕ 的 DID 估计量将是一致的吗？

(b) 如果可以利用面板数据，会有 ϕ 的一致估计吗？请解释你的回答。

22-5　当工具集合被推广到包括 lnwg、kids、age、agesq 以及 disab 的三个滞后项时，同时 1982～1988 年的 7 年数据可用于估计式(22.22)，利用齐利亚克(Ziliak，1997)的小时与工资数据，尽可能重新制作表 22.2 的大部分内容，并进行适当讨论。

23
非线性面板模型

23.1 引 论

本章将第 21 章和第 22 章的线性模型面板数据方法推广到由第 14 章至第 20 章所阐述的非线性回归模型。我们关注短面板以及含有时常值特定个体效应的模型，而时常值特定个体效应可能是固定的或是随机的。本章既考察静态模型，又考察动态模型。

就具有特定个体效应的非线性模型而言，不存在任何一种万能描述。倘若特定个体效应是固定的且面板数据是短的，则仅对非线性模型的一个子集才可能获得斜率参数的一致估计。不过，若特定个体效应是纯随机的，则对更广泛模型来说，都可能获得一致估计。

23.2 节阐述对特殊模型来说可能实施也可能不实施的一般方法。23.3 节给出具有乘法特定个体效应的非线性模型的一个应用。23.4～23.7 节对一些重要非线性模型，诸如离散数据、选择模型、过渡数据以及计数数据模型进行专门研究。23.8 节将对半参数估计提供一个综述。

23.2 一般结果

本节提供如何将线性模型的一些方法加以推广的一般方法。首先，以对条件均值模型与参数模型进行区别的方式来阐述几种模型，包括固定效应模型、随机效应模型以及混合模型。然后，讨论估计这些模型的方法，以及获得面板稳健的标准误差。而对特定非线性面板模型的进一步研究由下面一些小节给出。

23.2.1 *特定个体效应模型*

线性特定个体效应模型(参见 21.2.1 节)设定如下：因变量 y_{it} 依赖于时常值特定个体效应 α_i，以及通常回归元 $\mathbf{x}_{it}$ 与回归参数 $\boldsymbol{\beta}$。该模型可写成 $y_{it}=\alpha_i+\mathbf{x}_{it}'\boldsymbol{\beta}+u_{it}$，其中，$u_{it}$ 表示误差项。

对于非线性模型，诸如 logit 与泊松模型，缺乏引入可加误差 u_{it} 的动机。不过，一种更自然的方法是，直接对条件密度或条件均值进行建模，在线性情况下，条件

均值被写成 $\mathrm{E}[y_{it}|\alpha_i,\mathbf{x}_{it}]=\alpha_i+\mathbf{x}_{it}'\boldsymbol{\beta}$ 。

参数模型

对许多非线性模型来说，包括最著名的二值、多项式以及第 14～16 章给出的删失结果模型，完全参数方法是一种共同建模方式。

标准横截面模型是单指标函数，或具有附加标度参数的单指标模型。后面小节中将阐述的参数特定个体效应模型（**parametric individual-specific effects models**）是将条件密度设定成：

$$f(y_{it}|\alpha_i,\mathbf{x}_{it})=f(y_{it},\alpha_i+\mathbf{x}_{it}'\boldsymbol{\beta},\gamma) \tag{23.1}$$

其中，γ 表示附加参数，比如方差参数。该模型关于回归元 $\mathbf{x}_{it}$ 与个体效应 α_i 是单指标模型。

通常假设是 $y_{it}|\mathbf{x}_{it},\alpha_i$ 对不同 i 和不同 t 都是独立的。给定 i，该假设可被放松成在不同 t 上是相关的（参见 23.2.6 节）。

条件均值模型

一种相当一般的关于条件均值的非线性模型是含有时常值特定个体效应的模型，即：

$$\mathrm{E}[y_{it}|\alpha_i,\mathbf{x}_{it}]=g(\alpha_i,\mathbf{x}_{it},\boldsymbol{\beta}),\quad i=1,\cdots,N,\ t=1,\cdots,T \tag{23.2}$$

其中，$g(\cdot)$ 为已知函数。有三种普遍设定，第一种是可加特定个体效应模型（**additive individual-specific effects model**）：

$$g(\alpha_i,\mathbf{x}_{it},\boldsymbol{\beta})=\alpha_i+g(\mathbf{x}_{it},\boldsymbol{\beta}) \tag{23.3}$$

第二种是乘法特定个体效应模型（**multiplicative individual-specific effects model**）：

$$g(\alpha_i,\mathbf{x}_{it},\boldsymbol{\beta})=\alpha_i g(\mathbf{x}_{it},\boldsymbol{\beta}) \tag{23.4}$$

第三种是单指标特定个体效应模型（**singe-index individual-specific effects model**）：

$$g(\alpha_i,\mathbf{x}_{it},\boldsymbol{\beta})=g(\alpha_i+\mathbf{x}_{it}'\boldsymbol{\beta}) \tag{23.5}$$

在每一种形式里，函数 $g(\cdot)$ 都是设定的。回归元 $\mathbf{x}_{it}$ 可能是时变的或时常值的，并且可能包括时间虚拟变量。

当隐含的假设具有线性回归时，可加效应模型适合 y_{it} 的范围为无界的情况。乘法效应模型适合 y_{it} 为非负无界的情况，比如计数数据，在此情况下 $\alpha_i>0$ 且 $g(\cdot)>0$。单指标模型是 probit 模型的一个自然起点，例如 $g(\alpha_i+\mathbf{x}_{it}'\boldsymbol{\beta})=\Phi(\alpha_i+\mathbf{x}_{it}'\boldsymbol{\beta})$，其中，$\Phi(\cdot)$ 表示标准正态 cdf。当 $g(\cdot)$ 是恒等函数时，单指标模型就简化成可加模型。当 $g(\cdot)$ 是指数函数时，单指标模型便简化为乘法模型，从而 $\exp(\alpha_i+\mathbf{x}_{it}'\boldsymbol{\beta})=\exp(\alpha_i)\exp(\mathbf{x}_{it}'\boldsymbol{\beta})$。

矩条件（23.2）只以当前时期为条件，且假定回归元是同时期外生的（**contemporaneously exogenous**）（参见 22.2.4 节）。剔除特定个体效应需要较强的外生假设。若：

$$\mathrm{E}[y_{it}|\alpha_i,\mathbf{x}_{i1},\cdots,\mathbf{x}_{it}]=g(\alpha_i,\mathbf{x}_{it},\boldsymbol{\beta}) \tag{23.6}$$

则回归元是弱外生的，若：

$$E[y_{it}|\alpha_i, \mathbf{x}_{i1}, \cdots, \mathbf{x}_{iT}] = g(\alpha_i, \mathbf{x}_{it}, \boldsymbol{\beta}) \tag{23.7}$$

则回归元是强外生的或严格外生的。

含有可加效应的非线性模型相对而言增加了几个新困难。尤其是，如果面板模型是 $y_{it}=\alpha_i+g(x_{it},\boldsymbol{\beta})+u_{it}$，那么第 21 章与第 22 章的方法，包括通过非线性最小二乘法与工具变量而不是线性最小二乘法与工具变量进行估计，需做某种修改才能用于估计。

本章关注含有非线性特定个体效应的模型，比如式(23.4)与式(23.5)。这些效应将被处理成固定效应或随机效应。

23.2.2 *固定效应模型*

固定效应模型将特定个体效应处理成不可观测随机变量，该不可观测随机变量可能与回归元 $\mathbf{x}_{it}$ 相关。在短面板数据模型中，一般地讲，对固定效应 $\alpha_1, \cdots, \alpha_N$ 与模型其他参数 $\boldsymbol{\beta}$，可能还有 γ 进行联合估计，得到所有参数的非一致估计。不过，在一些特殊背景下，提出了一系列剔除固定效应的方法，使对模型其他参数得到一致估计成为可能。

附带参数问题

内曼和斯科特(Neyman and Scott, 1948)曾经考察，当某些参数对所有观测值来说是共同的，额外参数却有无限多个，其中每一个参数仅仅依赖于有限个观测值情形的推断问题。公共参数(**common parameters**)是人们内在关注的焦点，而后一种参数称为非主要参数[1](**incidental parameters**)。

这里 $\boldsymbol{\beta}$ 与 γ 均是公共参数，但 $\alpha_1, \cdots, \alpha_N$ 是非主要参数，倘若面板数据是短的，则每个 α_i 依赖于固定的 T 个观测值，而当 $N\to\infty$时，存在无限多个 α_i。由于仅有 T 个观测值用于估计每个参数，当 $N\to\infty$时，非主要参数被非一致地估计出来。非主要参数问题意指，此种情况污染了公共参数的估计。一般地讲，尽管公共参数是有限的，且可利用 $NT\to\infty$个观测值加以估计，但公共参数还是被非一致地估计出。

对起因于非主要参数污染的一种简单解释是，假定 $y_{it}\sim\mathcal{N}[\alpha_i,\sigma^2]$。运用极大似然法进行估计，得出 $\hat{\alpha}_i=\bar{y}_i, i=1,\cdots,N$，并且 $\hat{\sigma}^2=(NT)^{-1}\sum_i\sum_t(y_{it}-\bar{y}_i)^2$。于是，当 $N\to\infty$时，在固定 T 时短面板背景下，$E[\hat{\sigma}^2]=\sigma^2(T-1)/T$，故 $\hat{\sigma}^2$ 是 σ^2 的非一致估计值。当 $T=2$ 时，就 $\hat{\sigma}^2\xrightarrow{p}0.5\sigma^2$ 而言，这种非一致性可以非常大。

一般地讲，若存在非主要参数问题，则需要另一种估计方法，该方法首先剔除非主要参数。对于某些流行的模型来说，最著名的是面板 probit 模型，非主要参数问题没有解。甚至就存在一致估计的方法而言，这些方法倾向于使模型变成特定的，正如兰开斯特(Lancaster, 2000)所强调的。不存在统一求解非主要参数问题的方法。

[1] 又称为偶发参数。——译者注

条件似然

一个统计量 t 称为参数充分统计量，如果给定 t 时样本分布不依赖 θ。对于特定个体效应面板模型，若冗余参数 α_i 存在一个充分统计量，则通过以该充分统计量为条件就能剔除冗余参数 α_i。所得到的条件密度仅仅依赖于公共参数，从而得出一致估计。

设 $\mathbf{y}_i=[y_{i1},\cdots,y_{iT}]'$表示个体 i 在所有 T 时期的因变量 $T\times 1$ 维向量，设$\mathbf{X}_i=[\mathbf{x}_{i1},\cdots,\mathbf{x}_{iT}]'$表示相对应的回归元 $T\times K$ 阶矩阵。对于静态模型，$\mathbf{y}_i$ 具有密度：

$$f(\mathbf{y}_i\mid\mathbf{X}_i,\alpha_i,\boldsymbol{\beta},\boldsymbol{\gamma})=\prod_{t=1}^{T}f(y_{it}\mid\mathbf{x}_{it},\alpha_i,\boldsymbol{\beta},\boldsymbol{\gamma})\tag{23.8}$$

在短面板情况下，建立在该密度基础上的极大似然估计通常得出 $\boldsymbol{\beta}$ 的非一致估计，原因在于出现了非主要参数。

假定 α_i 存在一个充分统计量(**sufficient statistic**)$\mathbf{s}_i$。于是，除通常以回归元为条件以外，还以充分统计量 $\mathbf{s}_i$ 为条件，就得到条件密度：

$$f(\mathbf{y}_i\mid\mathbf{X}_i,\alpha_i,\boldsymbol{\beta},\boldsymbol{\gamma},\mathbf{s}_i)=f(\mathbf{y}_i\mid\mathbf{X}_i,\boldsymbol{\beta},\boldsymbol{\gamma},\mathbf{s}_i)\tag{23.9}$$

因此，α_i 被剔除。例如，对于线性回归模型，在正态条件下，$\mathbf{s}_i=\bar{y}_i$(参见 21.6.3 节)。从而，条件 MLE 对条件对数似然：

$$\ln \mathrm{L}_{\mathrm{COND}}(\boldsymbol{\beta},\boldsymbol{\gamma})=\sum_{i=1}^{N}\ln f(\mathbf{y}_i\mid\mathbf{X}_i,\boldsymbol{\beta},\boldsymbol{\gamma},\mathbf{s}_i)\tag{23.10}$$

求极大值。这里增加一个定语“条件”意指以 $\mathbf{s}_i$ 为条件，而不是以 $\mathbf{X}_i$ 为条件。

安德森(Andersen, 1970)对条件 MLE 进行了详细分析。他已经证明，若密度 $f(\mathbf{y}_i\mid\mathbf{X}_i,\alpha_i,\boldsymbol{\beta})$被正确设定，就条件对数似然而言，信息矩阵成立，则条件 MLE 是一致的。可是，一般地讲，由于条件 MLE 不必达到 Cramer-Rao 下界，所以出现效率损失。不过，对于正态分布与泊松分布来说，几乎没有效率损失。

需要适合充分统计量的方法确实存在。这种情况只针对少数几个模型，基本上是线性指数族的那些模型。安德森关注没有回归元的模型，并给出了正态、泊松、二项以及伽玛模型作为例子。一旦引进回归元，要找到适合的充分统计量甚至更加困难。麦卡拉和内尔德(McCullagh and Nelder, 1989)对此给出一种相当一般的讨论，而迪格尔等人(Diggle et al., 2002)将关注点限制在特殊化的具有标准连接函数[1](canonical link function)GLM。

就可利用充分统计量而言，重要例子是，正态性条件下的线性模型(参见 21.6.2 节)、二值数据的 logit 模型(尽管不是 probit 模型)、单参数伽玛(包括指数)、关于计数数据的特定参数化的泊松与负二项模型。

均值差分变换

对于含有可加或乘法效应的某些条件均值模型来说，个体效应可通过运用适当差分变换剔除。这样做就会产生能用于矩方法或 GMM 估计的矩条件，正如

---

〔1〕这里译者将其译为标准连接函数，此术语是广义线性模型中特有的，不要将它与“copulae”混淆。——译者注

23.2.6 节所述。

均值差分变换是 21.2.2 节给出的线性模型的组内变换的一种推广，那里通过减去特定个体均值来剔除。它要求强外生回归元，参见式(23.7)。

对于式(23.3)定义的可加效应模型，含有强外生回归元，从而：

$$\mathrm{E}[(y_{it}-\bar{y}_i)-(g(\mathbf{x}_{it}'\boldsymbol{\beta})-\bar{g}_i(\boldsymbol{\beta}))\,|\,\mathbf{x}_{i1},\cdots,\mathbf{x}_{iT}]=0 \qquad (23.11)$$

其中，$\bar{g}_i(\boldsymbol{\beta})=T^{-1}\sum_{t=1}^{T}g(\mathbf{x}_{it}'\boldsymbol{\beta})$，该结果使用了 $\mathrm{E}[\bar{y}_i\,|\,\mathbf{x}_{i1},\cdots,\mathbf{x}_{iT}]=\alpha_i+\bar{g}_i(\boldsymbol{\beta})$。就线性模型(23.11)而言，可进一步简化成 $g(\mathbf{x}_{it}'\boldsymbol{\beta})-\bar{g}_i(\boldsymbol{\beta})=(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)'\boldsymbol{\beta}$。

对于式(23.4)定义的乘法效应模型，经过某些代数运算，得到：

$$\mathrm{E}\left[y_{it}-\frac{g(\mathbf{x}_{it}'\boldsymbol{\beta})}{\bar{g}_i(\boldsymbol{\beta})}\times\bar{y}_i\,|\,\mathbf{x}_{i1},\cdots,\mathbf{x}_{iT}\right]=0 \qquad (23.12)$$

这里用到了 $\mathrm{E}[\bar{y}_i\,|\,\mathbf{x}_{i1},\cdots,\mathbf{x}_{iT}]=\alpha_i\bar{g}_i(\boldsymbol{\beta})$。为了简单起见，我们将此称为均值差分变换，尽管严格地讲，它是一个拟差分(**quasi-difference**)。它也称为(条件)均值标度变换，因为等价地有：

$$\mathrm{E}\left[y_{it}-\frac{\bar{y}_i}{\bar{g}_i(\boldsymbol{\beta})}g(\mathbf{x}_{it}'\boldsymbol{\beta})\,|\,\mathbf{x}_{i1},\cdots,\mathbf{x}_{iT}\right]=0$$

一阶差分变换

一阶差分变换(**first-differences transformation**)是 21.2.2 节给出的线性模型的一阶差分变换的推广，那里通过减去滞后一期模型来剔除 α_i。我们假定回归元都是弱外生的[参见式(23.6)]。

对于可加效应模型，有：

$$\mathrm{E}[(y_{it}-y_{i,t-1})-(g(\mathbf{x}_{it}'\boldsymbol{\beta})-g(\mathbf{x}_{i,t-1}'\boldsymbol{\beta}))\,|\,\mathbf{x}_{i1},\cdots,\mathbf{x}_{i,t-1}]=0 \qquad (23.13)$$

这里用到了 $\mathrm{E}[y_{i,t-1}\,|\,\mathbf{x}_{i1},\cdots,\mathbf{x}_{i,t-1}]=\alpha_i+g(\mathbf{x}_{i,t-1}'\boldsymbol{\beta})$。

对于式(23.4)定义的乘法效应模型，有：

$$\mathrm{E}\left[y_{it}-\frac{g(\mathbf{x}_{it}'\boldsymbol{\beta})}{g(\mathbf{x}_{i,t-1}'\boldsymbol{\beta})}\times y_{i,t-1}\,|\,\mathbf{x}_{i1},\cdots,\mathbf{x}_{i,t-1}\right]=0 \qquad (23.14)$$

其中，我们用到了 $\mathrm{E}[y_{i,t-1}\,|\,\mathbf{x}_{i1},\cdots,\mathbf{x}_{i,t-1}]=\alpha_i g(\mathbf{x}_{i,t-1}'\boldsymbol{\beta})$。为了简单起见，我们称为一阶差分变换，尽管严格地讲，它是一个拟差分(**quasi-difference**)。仅以直到时期 t 为条件的一阶差分变换，依赖于弱假设。它使得对 22.5 节推广到非线性模型的动态模型进行估计成为可能。对于动态乘法效应模型，伍德里奇(Wooldridge, 1997)与张伯伦(Chamberlain, 1992)实际上提出使用式(23.14)的一种变形，即：

$$\mathrm{E}\left[\frac{g(\mathbf{x}_{i,t-1}'\boldsymbol{\beta})}{g(\mathbf{x}_{it}'\boldsymbol{\beta})}y_{it}-y_{i,t-1}\,|\,\mathbf{x}_{i1},\cdots,\mathbf{x}_{i,t-1}\right]=0 \qquad (23.15)$$

虚拟变量模型估计

如果忽略非主要参数问题，人们就能企图估计所有参数，包括特定个体效应。引入 N 个虚拟变量 $d_{j,it}$ 集合，当 $i=j$ 时，$d_{j,it}=1$，否则 $d_{j,it}=0$，然后联合估计特定个体效应参数 $\alpha_1,\cdots,\alpha_N$ 以及模型的其他参数。

尽管由大 N 导致了相当多的参数，从计算上看，此估计量是可行的，但得到的 $\boldsymbol{\beta}$ 以及估计值 $\boldsymbol{\gamma}$ 一般可能是非一致的。这里，我们只考察参数模型，可是对条件均值模型来说，类似情况也成立。

因而，考察由式(23.1)定义的参数形式特定个体效应模型。于是，运用对整个对数似然函数：

$$\ln \mathrm{L}_{\mathrm{FE}}(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\alpha}) = \sum_{i=1}^{N}\sum_{t=1}^{T}\ln f(y_{it},\mathbf{d}_{it}'\boldsymbol{\alpha} + \mathbf{x}_{it}'\boldsymbol{\beta},\boldsymbol{\gamma}) \tag{23.16}$$

求极大值的方法得出 $\boldsymbol{\beta}$、$\boldsymbol{\gamma}$ 以及 $\boldsymbol{\alpha} = [\alpha_1 \cdots \alpha_N]'$ 的极大似然估计值，其中，$\mathbf{d}_{it} = [d_{1,it} \cdots d_{N,it}]'$。关于 $\boldsymbol{\delta}=[\boldsymbol{\beta}' \ \boldsymbol{\gamma}']'$ 与 $\boldsymbol{\alpha}$ 的一阶条件是：

$$\sum_{i=1}^{N}\sum_{t=1}^{T}\partial \ln f(y_{it},\mathbf{d}_{it}'\boldsymbol{\alpha} + \mathbf{x}_{it}'\boldsymbol{\beta},\boldsymbol{\gamma})/\partial\boldsymbol{\delta} = \mathbf{0}$$

$$\sum_{t=1}^{T}\partial \ln f(y_{it},\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta},\boldsymbol{\gamma})/\partial\alpha_i = 0, \quad i = 1,\cdots,N$$

尽管参数个数 N 加上 $\boldsymbol{\delta}$ 的维数会更大，但仍能直接计算该估计量。如同格林(Greene, 2004b)详细讨论的，海赛矩阵的逆很容易通过对 $\boldsymbol{\delta}$ 与 $\boldsymbol{\alpha}$ 进行分块，并运用标准分块逆公式得到，对于 $j \neq i$，利用 $\partial \ln \mathrm{L}(\boldsymbol{\delta},\boldsymbol{\alpha})/\partial\alpha_i\partial\alpha_j = 0$ 加以简化，故对应于$(\boldsymbol{\alpha},\boldsymbol{\alpha})$的 $N \times N$ 块逆容易求出。

存在两种特殊情况没有非主要参数问题。第一种情况是，若 $y_{it} \sim \mathcal{N}[\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}, \sigma^2]$，则由 21.6.4 节知，$\boldsymbol{\beta}$ 的极大似然估计是组内估计量，甚至对于有限 T 来说，关于 $\boldsymbol{\beta}$ 是一致的。这里的非主要参数问题起因于 σ^2 的而不是 $\boldsymbol{\beta}$ 的估计。第二种情况是，类似地，对于 $y_{it} \sim \mathcal{P}[\exp(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta})]$，估计 $\boldsymbol{\beta}$ 时，不存在非主要参数问题(参见 23.7.3 节)。

不过，一般地讲，存在非主要参数问题。关于 α_i 的推导仅仅涉及 T 个观测值，而不是所有 NT 个观测值。在短面板数据中，这时常产生 $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ 与 $\hat{\boldsymbol{\gamma}}_{\mathrm{ML}}$ 的非一致性。在不太短的面板数据比如 $T=10$ 或 $T=20$ 的情况下，此非一致性可能是适度的。格林(Greene, 2004a)的模拟研究表明，偏倚的特性及范围会随着所探讨的特殊非线性模型而出现相当大的变化。在存在固定效应的条件下，发展稳健方法是研究领域中一个活跃专题，尽管短面板数据仍然有非一致性。

23.2.3 *随机效应模型*

随机效应模型是特定个体效应 α_i 处理成服从设定分布的随机变量，并通过对该分布进行积分去掉 α_i。随机效应通常应用于参数模型。

参数模型

假定第 i 个观测值 $\mathbf{y}_i$ 具有式(23.8)给出的无条件联合密度 $f(\mathbf{y}_i|\mathbf{X}_i,\alpha_i,\boldsymbol{\beta},\boldsymbol{\gamma})$，并且其随机效应具有密度：

$$\alpha_i \sim g(\alpha_i|\boldsymbol{\eta}) \tag{23.17}$$

其中，$g(\alpha_i|\boldsymbol{\eta})$不依赖于可观测值。于是，第 i 个观测值的无条件联合密度是：

$$f(\mathbf{y}_i \mid \mathbf{X}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\eta}) = \int \Big[\prod_{t=1}^{T} f(y_{it} \mid \mathbf{x}_{it}, \alpha_i, \boldsymbol{\beta}, \boldsymbol{\gamma}) \Big] g(\alpha_i \mid \boldsymbol{\eta}) d\alpha_i \tag{23.18}$$

其中，我们用无条件意指不再以 α_i 为条件。$\boldsymbol{\beta}$、$\boldsymbol{\gamma}$、$\boldsymbol{\eta}$的随机效应极大似然估计是对对数似然

$$\ln \mathrm{L}_{\mathrm{RE}}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\eta}) = \sum_{t=1}^{N} \ln\Big(\int \Big[\prod_{t=1}^{T} f(y_{it} \mid \mathbf{x}_{it}, \alpha_i, \boldsymbol{\beta}) \Big] g(\alpha_i \mid \boldsymbol{\gamma}) d\alpha_i \Big) \tag{23.19}$$

求极大值。

在一些情况下，从根本上讲，当$\prod_t f(y_{it} \mid \alpha_i)$与 $g(\alpha_i)$是共轭对（参见表 13.2）时，则积分可能是解析表达式。这样的例子包括得到正态结果的正态—正态线性回归，以及得出负二项式结果的泊松--伽玛计数数据回归。

在大多数情况下，没有解析结果可以利用，但数值方法或基于模拟方法可能更适合，因为积分仅仅是一维的。在不存在个体效应时，通常方法是选择 $f(y_{it})$作为被认为最佳拟合数据的密度，然后设 $g(\alpha_i)$是正态密度。于是，积分是关于正态随机变量的一元积分。对于小 T 来说，该积分通过高斯—埃尔米特求积法得到很好近似，它是借助于加权和对正态密度进行逼近的。巴特勒和莫菲特（Butler nad Moffitt, 1982）对随机效应 probit 模型给出了详细阐述。斯克龙达尔和拉贝·哈斯克特（Skrondal and Rabe-Hasketh, 2004）则运用求积法。一种可能的选择方式是从 $g(\alpha_i)$中重复采样以作为模拟极大似然估计的基础（参见 12.4.2 节）。

上述讨论假定给定 i 对不同 t 具有独立性。不过，倘若对不同 i 来说，y_{it}与 y_{is}是相关的，则更有效方法是，用 $f(\mathbf{y}_i \mid \mathbf{X}_i, \alpha_i, \boldsymbol{\beta}, \boldsymbol{\gamma})$代替式（23.18）与式（23.19）中的$\prod_t f(y_{it} \mid \mathbf{x}_{it}, \alpha_i, \boldsymbol{\beta}, \boldsymbol{\gamma})$。

随机效应模型

类似于 22.8 节的线性情况，很明显，可将随机效应方法推广到含有随机斜率与随机截距的随机系数模型上。

一个正常模型是具有条件密度 $f(y_{it}, \mathbf{x}'_{it}(\boldsymbol{\beta} + \boldsymbol{\alpha}_i), \boldsymbol{\gamma})$或条件均值 $g(y_{it}, \mathbf{x}'_{it}(\boldsymbol{\beta} + \boldsymbol{\alpha}_i))$的单指标模型，并且关于纯量 α_i 的一元积分将变成关于向量 $\boldsymbol{\alpha}_i$的多元积分，通常假设 α_i 服从正态分布。

相关随机效应模型

随机效应模型的一个重要弱点是，它做出如下假设：随机效应与回归元是独立的。为了克服这个局限性，张伯伦（Chamberlain, 1980, 1982）提出相关的随机效应模型，其有关背景讨论参见 21.4.4 节，该模型设定如下：

$$\alpha_i = \mathbf{x}'_{1i}\boldsymbol{\pi}_1 + \cdots + \mathbf{x}'_{Ti}\boldsymbol{\pi}_T + \xi_i \tag{23.20}$$

那么，上述似然函数是对 $\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\pi}$ 以及密度的 ξ 参数求极大值。与线性模型不同，这个模型会得出不同于利用芒德拉克（Mundlak, 1978）的较简单设定而得到的估计量：

$$\alpha_i = \bar{\mathbf{x}}'_i \boldsymbol{\pi} + \xi_i \tag{23.21}$$

可将式(23.20)看成分层模型的一个例子。更一般的分层模型同样允许出现随机斜率,并利用经典方法或贝叶斯方法加以估计。22.8节已经详细阐述了线性模型。

有限混合模型

有限混合模型(参见18.5.1节)提供了不可观测特定个体效应的一种可供选择模型。若存在m种个体类型或潜在类别,对于第j个类型$\alpha_i=\alpha_j$,则式(23.18)变成:

$$f(\mathbf{y}_i \mid \mathbf{X}_i,\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\pi}) = \sum_{j=1}^{m}\left[\prod_{t=1}^{T} f(y_{it} \mid \mathbf{x}_{it},\alpha_j,\boldsymbol{\beta},\boldsymbol{\gamma})\right]\pi_j$$

该模型最常用于面板持续期限模型(参见18.5.2节)。

23.2.4 *混合模型*

混合模型并没有对特定个体效应以显性方式进行建模。它是将线性混合回归推广到非线性模型上。

条件均值模型

对于条件均值模型来说,混合模型是:

$$\mathrm{E}[y_{it} \mid \mathbf{x}_{it}]=g(\mathbf{x}_{it},\boldsymbol{\beta}) \tag{23.22}$$

其中,$g(\cdot)$为设定函数。

模型(23.22)能直接通过NLS进行估计,其推断建立在面板稳健标准误差的基础上,这样控制了条件异方差性与y_{it}及y_{is}之间的条件相关。一种更为有效的估计,可通过对异质性与相关进行建模。其详细内容将由23.2.6节给出。

混合模型与随机效应模型

忽略特定个体随机效应的代价是什么呢?

当$\mathrm{E}[\alpha_i \mid \mathbf{x}_{it}]=0$时,可加效应模型$\mathrm{E}[y_{it} \mid \alpha_i,\mathbf{x}_{it}]=\alpha_i+g(\mathbf{x}_{it},\boldsymbol{\beta})$会得出式(23.22)。当$\mathrm{E}[\alpha_i \mid \mathbf{x}_{it}]=1$时,乘法效应模型$\mathrm{E}[y_{it} \mid \alpha_i,\mathbf{x}_{it}]=\alpha_i g(\mathbf{x}_{it},\boldsymbol{\beta})$蕴含式(23.22)。因此,若效应是可加的或乘法形式,且可以使用这些模型α_i均值的标准正规化,则混合效应模型将得到随机效应模型$\boldsymbol{\beta}$的一致估计。

否则,混合模型将不可能得出与特定个体随机效应模型一样的参数估计。例如,考察满足$\mathrm{E}[y_{it} \mid \alpha_i,\mathbf{x}_{it}]=\Phi(\alpha_i+\mathbf{x}_{it}'\boldsymbol{\beta})$的probit随机变量模型,其中,$\alpha_i\sim\mathcal{N}[0,\sigma_\alpha^2]$。那么,可以证明,$\mathrm{E}[y_{it} \mid \mathbf{x}_{it}]=\Phi(\mathbf{x}_{it}'\boldsymbol{\beta}/\sqrt{1+\sigma_\alpha^2})$,这不同于正常的混合probit模型$\mathrm{E}[y_{it} \mid \mathbf{x}_{it}]=\Phi(\mathbf{x}_{it}'\boldsymbol{\beta})$。与第21章的线性模型不同,倘若真实模型具有特定个体随机效应,并忽略这个随机效应,则会得出$\boldsymbol{\beta}$的非一致参数估计。

统计学文献对于广义线性模型比如二值数据与计数数据的面板形式,广泛运用混合模型方法。所得到的参数估计称为总体平均,因为已剔除了随机效应。这种方法称为边际分析,这是因为$\mathrm{E}[y_{it} \mid \mathbf{x}_{it}]$作为针对随机效应的边际模型。

参数模型

就混合参数模型而言,其起点通常是:

$$f(y_{it} \mid \mathbf{x}_{it})=f(y_{it},\mathbf{x}_{it}'\boldsymbol{\beta},\boldsymbol{\gamma}) \tag{23.23}$$

其中，$f(\cdot)$设定函数。该模型利用极大似然法加以估计，其推断建立在控制条件异方差性与相关性的面板稳健标准误差基础上。

一般地讲，$\boldsymbol{\beta}$ 及 γ 的混合参数模型不可能与源于随机效应参数模型的那些值相一致。对此推理类似于条件均值的那种情况。

23.2.5 固定效应与随机效应

若引入特定个体效应且与回归元相关，则随机效应与混合模型估计量是非一致的这个基本结果对非线性模型来说仍然成立。考虑到稳健性，人们更愿意运用固定效应模型，尽管估计时会存在对有效性损失的权衡。豪斯曼检验能用作检验是否需要固定效应模型，倘若可能得到固定效应模型的一致估计值。

有关线性模型的固定效应模型与随机效应模型的其他比较，需要做某种修改才能用于非线性模型。

因为非主要参数问题，所以不是所有具有固定效应的非线性模型都允许得出一致参数估计值。因而，固定效应建模并不总是可行的。

若非线性固定效应模型可能得出一致估计，则与线性情况不同，时常值回归元的系数就是可识别的。为了理解这一点，考察可加效应模型的均值差分变换。对于线性模型 $\mathrm{E}[(y_{it}-\bar{y}_i)-(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)'\boldsymbol{\beta}\mid\mathbf{x}_{i1},\cdots,\mathbf{x}_{iT}]=\mathbf{0}$，就时常值回归元的明显问题而言，考虑第 j 个回归元。更一般地讲，由式(23.11)知：

$$\mathrm{E}[(y_{it}-\bar{y}_i)-(g(\mathbf{x}_{it}'\boldsymbol{\beta})-\bar{g}_i(\boldsymbol{\beta}))\mid\mathbf{x}_{i1},\cdots,\mathbf{x}_{iT}]=\mathbf{0}$$

对非线性的 $g(\cdot)$来说，并不存在这类简化，除非 $\mathbf{x}_{it}$ 的全部 K 个成分都是时常值的。

在具有非可加效应的固定效应模型中，当回归元变动，不可能预测因变量的变化。就一般模型(23.2)而言，边际效应$\partial\mathrm{E}[y_{it}\mid\mathbf{x}_{it},\alpha_i,\boldsymbol{\beta}]/\partial\mathbf{x}_{it}=\partial g(\mathbf{x}_{it},\alpha_i,\boldsymbol{\beta})/\partial\mathbf{x}_{it}$依赖于$\alpha_i$。

可以测量两种特殊情况下的边际效应。对于可加效应(参见 23.3 节)，其边际效应是$\partial g(\mathbf{x}_{it},\boldsymbol{\beta})/\partial\mathbf{x}_{it}$，该值不依赖于 α_i。对于乘法效应模型(参见 23.4 节)，其边际效应是 $\alpha_i\partial g(\mathbf{x}_{it},\boldsymbol{\beta})/\partial\mathbf{x}_{it}$。于是，对不同回归元变动时，可能测量出边际效应的相对变化。尤其是，假如 $\mathrm{E}[y_{it}\mid\mathbf{x}_{it},\alpha_i,\boldsymbol{\beta}]=\alpha_i\exp(\mathbf{x}_{it}'\boldsymbol{\beta})$，则$(\partial\mathrm{E}[y_{it}]/\partial x_{itj})/(\partial\mathrm{E}[y_{it}]/\partial x_{itk})=\beta_j/\beta_k$。

23.2.6 估计与面板稳健统计推断

上述分析着重研究剔除非主要参数 α_i 的问题。现在，我们阐述当去掉含有特定个体效应模型的 α_i 时模型的参数估计。

我们假定短面板数据，并对不同 i 来说观测值具有独立性。因变量 y_{it} 可能是条件异方差的，且对给定 i 的不同 t 来说是条件相关的。这种情况类似于 21.2.3 节，只是用非线性估计量代替较简单的线性最小二乘法估计量。标准的统计输出忽略了这种复杂性，从而导致推断无效。下面内容阐述参数估计方差矩阵的稳健面板估计表达式。作为一种可供选择方式，能使用面板自助法(参见 11.6.2 节)。

广义矩方法估计

对于建立在条件均值基础上的模型，面板广义矩方法是合适的。其关键是对矩条件进行设定，这里的矩条件是广义矩方法估计的基础。沿着 22.2.1 节的线索，一个正常起点是：

$$\mathrm{E}[\mathbf{Z}_i'\mathbf{u}_i(\boldsymbol{\theta})]=\mathbf{0},\quad i=1,\cdots,N \tag{23.24}$$

其中，$\mathbf{Z}_i$ 表示 $T\times r$ 阶矩阵，该矩阵依赖于回归元，$\mathbf{u}_i(\boldsymbol{\theta})$ 表示 $T\times 1$ 维残差向量，$\boldsymbol{\theta}$ 表示 $q\times 1$ 维参数向量 $\boldsymbol{\theta}$。各种不同面板模型会导致对 $\mathbf{u}_i$ 与 $\mathbf{Z}_i$ 的不同设定。一个例子将在下面给出。对第 22 章的一个重要背离是，残差 $\mathbf{u}_i(\boldsymbol{\theta})$ 关于 $\boldsymbol{\theta}$ 是非线性的。

当 $r=q$ 时，存在与参数同样多的矩条件用于估计，我们运用面板矩方法估计量 $\hat{\boldsymbol{\theta}}_{\mathrm{MM}}$，它是

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{u}_i(\hat{\boldsymbol{\theta}})=\mathbf{0} \tag{23.25}$$

的解。利用 6.10.3 节关于非线性系统估计的结果，得到该估计量服从渐近正态分布，其方差矩阵可通过

$$\hat{\mathrm{V}}[\hat{\boldsymbol{\theta}}]=\left[\sum_{i=1}^{N}\hat{\mathbf{D}}_i'\mathbf{Z}_i\right]^{-1}\sum_{i=1}^{N}\mathbf{Z}_i'\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'\mathbf{Z}_i\left[\sum_{i=1}^{N}\mathbf{Z}_i'\hat{\mathbf{D}}_i\right]^{-1} \tag{23.26}$$

得到一致估计，其中，$\hat{\mathbf{D}}_i=\partial\mathbf{u}_i/\partial\boldsymbol{\theta}'|_{\hat{\boldsymbol{\theta}}}$，$\hat{\mathbf{u}}_i=\mathbf{u}_i(\hat{\boldsymbol{\theta}})$。从而，得到短面板的稳健面板标准误差。

当 $r>q$ 时，必须用广义矩方法估计，我们运用面板广义矩方法估计量 $\hat{\boldsymbol{\theta}}_{\mathrm{GMM}}$，它极小化：

$$Q_N(\boldsymbol{\theta})=\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{u}_i(\boldsymbol{\theta})\right]'\mathbf{W}_N\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{u}_i(\boldsymbol{\theta})\right] \tag{23.27}$$

其中，$\mathbf{W}_N$ 表示 $r\times r$ 阶加权矩阵。该估计量的渐近方差矩阵能从 6.10.4 节给出非线性系统工具变量估计量的结果直接获得。已知矩条件(23.24)，最有效估计量使用 $\mathbf{W}_N=[N^{-1}\sum_i\mathbf{Z}_i'\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'\mathbf{Z}_i]^{-1}$。

更有效的估计量可能利用可供选择的矩条件。尤其是，若起点是特别的条件矩条件，则广义矩方法估计的最优无条件矩条件由 6.3.7 节给出。后面将给出的广义估计方程估计量就来自该方法。埃弗里、汉森和霍茨(Avery, Hansen, and Hotz, 1983)以及布赖通和莱奇纳(Breitung and Lechner, 1999)进行了更一般的研究。

广义矩方法例子

举一个特定例子，考察利用乘法固定效应模型的一阶差分变换。其起点是条件矩约束(23.14)。这会得到许多无条件矩条件，其中一个是：

$$\mathrm{E}\left[\mathbf{x}_{it}\left(y_{it}-\frac{g(\mathbf{x}_{it}'\boldsymbol{\beta})}{g(\mathbf{x}_{i,t-1}'\boldsymbol{\beta})}\times y_{i,t-1}\right)\right]=\mathbf{0},\quad t=1,\cdots,N$$

假定有$(T+1)$个时期的$(y_{it},\mathbf{x}_{it})$数据可以利用，由于进行一阶差分运算，故损失了最初时期数据。对 T 个时期数据叠放，得出式(23.24)，这里，$\mathbf{Z}_i'=[\mathbf{x}_{i1},\cdots,\mathbf{x}_{iT}]$，

$\mathbf{u}_i'=[u_{i1},\cdots,u_{iT}]$。其中，$u_{it}=y_{it}([g(\mathbf{x}_{it}'\boldsymbol{\beta})/g(\mathbf{x}_{i,t-1}'\boldsymbol{\beta})]y_{i,t-1}$。从而，$\mathbf{Z}_i'\mathbf{u}_i=\sum_t\mathbf{x}_{it}u_{it}$，因此，矩方法估计量 $\hat{\boldsymbol{\beta}}$ 是：

$$\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{x}_{it}\left[y_{it}-\frac{g(\mathbf{x}_{it}'\boldsymbol{\beta})}{g(\mathbf{x}_{i,t-1}'\boldsymbol{\beta})}y_{i,t-1}\right]=\mathbf{0}$$

的解。很明显，可以使用另外的矩条件，诸如 $\mathrm{E}[\mathbf{x}_{it-1}u_{it}]=\mathbf{0}$，进而得到一个过度识别模型，并通过广义矩方法加以估计。第 22 章线性模型对此进行了广泛讨论。

广义估计方程估计

条件均值的混合模型设定 $\mathrm{E}[y_{it}|\mathbf{x}_{it}]=g(\mathbf{x}_{it},\boldsymbol{\beta})$（参见 23.2.4 节）。该模型能通过已阐述的广义矩方法进行估计。这里我们还要进一步研究并考察有效广义矩方法估计。

对所有 T 个观测值叠放，得到条件矩条件：

$$\mathrm{E}[\mathbf{y}_i-\mathbf{g}_i(\boldsymbol{\beta})|\mathbf{X}_i]=\mathbf{0} \tag{23.28}$$

其中，$\mathbf{g}_i(\boldsymbol{\beta})=[g(\mathbf{x}_{i1},\boldsymbol{\beta}),\cdots,g(\mathbf{x}_{iT},\boldsymbol{\beta})]'$，而 $\mathbf{X}_i=[\mathbf{x}_{i1},\cdots,\mathbf{x}_{iT}]'$。于是，用于估计的最优无条件矩条件是：

$$\mathrm{E}\left[\frac{\partial\mathbf{g}_i'(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\{\mathrm{V}[\mathbf{y}_i|\mathbf{X}_i]\}^{-1}(\mathbf{y}_i-\mathbf{g}_i(\boldsymbol{\beta}))\right]=\mathbf{0} \tag{23.29}$$

运用 6.3.7 节给出的一般结果，得出一个结果。从而，得到广义估计方程估计量 $\hat{\boldsymbol{\beta}}_{\mathrm{GEE}}$，它是

$$\sum_{i=1}^{N}\frac{\partial\mathbf{g}_i'(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i-\mathbf{g}_i(\boldsymbol{\beta}))=\mathbf{0} \tag{23.30}$$

的解，其中，$\boldsymbol{\Sigma}_i$ 表示 $\mathrm{V}[\mathbf{y}_i|\mathbf{X}_i]$ 的实用方差矩阵。$\hat{\boldsymbol{\beta}}_{\mathrm{GEE}}$ 的渐近方差矩阵已由式(23.26)给出，$\hat{\mathbf{u}}_i=\mathbf{y}_i-\mathbf{g}_i(\hat{\boldsymbol{\beta}})$ 以及 $\mathbf{Z}_i=\partial\mathbf{g}_i'(\boldsymbol{\beta})/\partial\boldsymbol{\beta}|_{\hat{\beta}}\times\hat{\boldsymbol{\Sigma}}_i$。此方差估计是面板稳健的，并对 $\boldsymbol{\Sigma}_i$ 的错误设定也是稳健的。

归功于梁和塞格尔(Liang and Zeger,1986)的广义估计方程估计量，在广义线性模型的面板形式统计学文献中被广泛使用。各种不同的广义线性模型对应于不同的条件均值函数 $\mathbf{g}_i(\boldsymbol{\beta})$ 与实用方差矩阵 $\boldsymbol{\Sigma}_i$。

极大似然估计

对于基于似然的模型，其讨论起点是所有 T 个个体的联合密度，即 $f(\mathbf{y}_i|\mathbf{X}_i,\boldsymbol{\theta})$。对于混合参数模型，$\boldsymbol{\theta}'=[\boldsymbol{\beta}',\boldsymbol{\gamma}']$[参见式(23.23)]，而对于随机效应参数模型，$\boldsymbol{\theta}'=[\boldsymbol{\beta}',\boldsymbol{\gamma}',\boldsymbol{\eta}']$[参见式(23.18)]。

一种标准方法是设 $f(\mathbf{y}_i|\mathbf{X}_i,\boldsymbol{\theta})=\prod_{t=1}^{T}f(y_{it}|\mathbf{x}_{it},\boldsymbol{\theta})$，其中，$f(y_{it}|\mathbf{x}_{it},\boldsymbol{\theta})$ 表示第 (i,t) 个观测值的密度。给定 i 对不同 t 具有独立性的隐性假设通常是无保证的，尤其是不包括随机效应的混合模型，随机效应会允许出现不同时间上的某种相关。不过，倘若 $f(y_{it}|\mathbf{x}_{it},\boldsymbol{\theta})$ 被正确设定，即使 $f(\mathbf{y}_i|\mathbf{X}_i,\boldsymbol{\theta})$ 被错误设定，也能得到 $\boldsymbol{\theta}$ 的一致估计。于是，为了保证面板稳健标准误差，应该将三明治形式用作估计量方差矩阵。极大似然估计是严格的拟极大似然估计，5.7.5 节对此进行了详细讨论。更一般地讲，这种方法是有关聚集数据推断的一个例子(参见 24.5 节)。

利用允许对不同时间出现相关的 $f(\mathbf{y}_i|\mathbf{X}_i,\boldsymbol{\theta})$ 的更为丰富模型，获得更有效估计是可能的。可是，关于 $\mathbf{y}_i$ 的非正态多元分布是一个约束，或者难以继续研究。对于混合广义线性模型，却要使用广义估计方程估计量。

23.2.7 *动态模型*

具有特定个体效应的动态模型是人们着重关注的内容，因为这种模型使得人们能区分真实状态相关性与由不可观测异质性引起的伪相关性(参见 22.5.1 节)。

对于非线性模型，如何包括滞后因变量作为回归元并不总是明显的，因为对某些数据类型来说，不是总存在标准的纯时间序列模型。就泊松模型而言，23.7.4 节将阐述这一点。当做出合适设定后，标准固定效应估计量就变成非一致的，为了并入初始条件，需要随机效应估计量，如同线性面板模型那样。

混合模型

混合模型忽略了随机效应，并对通常横截面模型加以估计，该横截面模型的回归元现在包括滞后因变量。这再次与 23.2.4 节所讨论的内容有关。

固定效应模型

就固定效应模型而言，问题类似于 22.5 节所述的那些问题。现在回归元是弱外生的而不是强外生的。通常固定效应估计量都是非一致的。

对于含有可加效应或乘法效应的模型，当使用一阶差分变换(参见 23.2.2 节)以及用滞后因变量的高较阶作为工具时，可能得出一致估计。对于可加效应模型，这会得到 22.5.3 节给出的阿雷拉诺—邦德估计量的非线性形式。对于乘法效应模型，其一阶差分变换将由 23.7.4 节详述。对于含有固定效应的动态 logit 模型，参见 23.4.3 节。

参数随机效应模型

就参数随机效应模型而言，有关滞后因变量的初始条件会发挥作用，一般地讲，并不存在令人满意的处理，因此，在短面板情况下估计都是非一致的，当 T 增大时，其非一致性将变小。

考察一种最简单的情况，即仅有一个时期滞后项出现在模型中，因而回归元 $\mathbf{x}_{it}$ 变成回归元 $\mathbf{x}_{it}$ 与 y_{it-1}。随机效应密度(23.1)变为 $f(y_{it}|y_{it-1},\mathbf{x}_{it},\alpha_i,\boldsymbol{\delta})$，对于 $t=2,\cdots,T$。不过，由于 y_{i0} 是不可观测的，故并不能包括关于 y_{i1} 的类似模型。一种方法是，将 y_{i1} 处理成外生的，因此，我们仅对 $T-1$ 个观测值 $y_{it},\cdots,y_{i2}$ 的条件分布进行建模。一种可供选择的方法是提出一个静态模型，假定 y_{i1} 依赖于回归元 $\mathbf{x}_{i1}$ 且可能依赖于边际效应 α_i。从而，$\mathbf{y}_i$ 的联合条件密度是：

$$
\begin{aligned}
& f(\mathbf{y}_i|\mathbf{x}_{i1},\cdots,\mathbf{x}_{iT},\alpha_i,\boldsymbol{\delta},\boldsymbol{\delta}_1,\gamma) \\
&= \int\Big[\prod_{t=2}^{T} f(y_{it}|y_{it-1},\mathbf{x}_{it},\alpha_i,\boldsymbol{\delta})\Big]f_1(y_{i1}|\mathbf{x}_{i1},\alpha_i,\boldsymbol{\delta}_1)g(\alpha_i|\gamma)d\alpha_i
\end{aligned}
$$

而不是式(23.18)，其中，$f_1(y_{i1}|\mathbf{x}_{i1},\alpha_i,\boldsymbol{\delta}_1)$ 表示第一个观测值的假定密度。

纯时间序列分析中，初始条件当 $T\to\infty$ 时会变得渐近无关。不过在短面板中，当 T 是小的数值时，初始条件就极为重要，而且渐近特性使用了 $T\to\infty$。

23.2.8 *内生回归元*

对非线性模型的内生变量的处理，类似于第 22 章阐述的线性模型情况。

面板广义矩方法是一个通行的框架。对于适当定义残差 $\mathbf{u}_i(\boldsymbol{\theta})$ 与工具 $\mathbf{Z}_i$ 来说，起点是条件矩约束 $\mathrm{E}[\mathbf{u}_i(\boldsymbol{\theta})|\mathbf{Z}_i]=\mathbf{0}$。这就得出作为广义矩方法估计基础的无条件件矩(23.24)。备选工具可能包括除当前一个时期以外的其他时期外生回归元，如同 22.2 节与 22.4 节对线性模型的讨论。

23.3 非线性面板例子：专利与研发

我们运用源自霍尔、格里利谢斯和豪斯曼(Hall, Griliches, and Hausman, 1986)的 1975～1979 年 5 年期间 346 个厂商的每一年数据，对专利与研究之间的关系进行建模。其因变量 y_{it} 表示专利，即在最终被授予年份期间申请专利数。为简单起见，我们仅仅考虑一个解释变量 x_{it}，即指定年份的实际研发支出(以 1972 年美元计算)。

一个明显起点模型是，对数—对数模型，满足 $\mathrm{E}[\ln y_{it}|x_{it}]=\alpha_i+\beta\ln x_{it}$，从而 β 等于专利研发弹性。该模型并不能用于这里，因为相当多的观测值出现 $y_{it}=0$，而 $\ln 0$ 则没有定义。一种特定调整是，在取对数之前将 $y_{it}=0$ 重新记录成 $y_{it}=0.5$。

图 23.1 画出运用全部厂商所有年份的数据得到的调整后的 ln(专利)与 ln(R&D)图，以及拟合 OLS(估计斜率系数为 0.834)和非参数回归曲线。很明显，专利随着 R&D 支出而增长。面板数据分析，尤其是固定效应模型，能将这种关系分解为横截面成分与时间序列成分。可以发现，专利对不同观测值特别是对不同厂商来说变化很大，其均值为 36.3，标准差为 74.5，其变化范围就所有年份全部厂商而言为从 0 到 608。

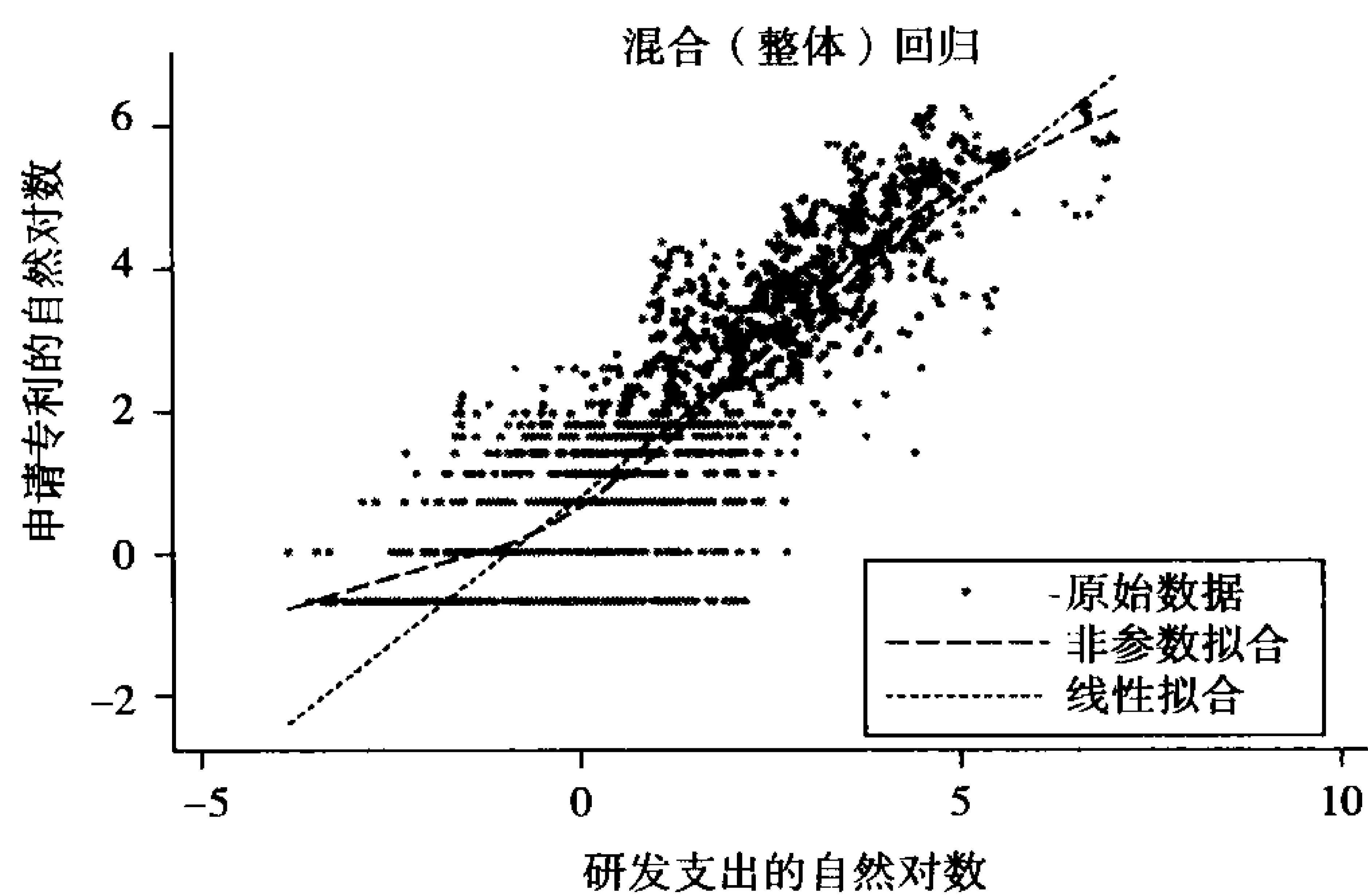


图 23.1 专利与研发支出：混合(整体)回归。关于 346 个厂商，1975～1979 年的 5 年期间每年申请专利的自然对数与研发支出的自然对数图形。零申请专利记为 0.5。

我们估计一个乘法特定个体效应模型，其条件均值为：

$$\mathrm{E}[y_{it}|x_{it},\alpha_i]=\alpha_i\exp(\beta\ln x_{it})=\exp(\gamma_i+\beta\ln x_{it}) \tag{23.31}$$

其中，$\gamma_i = \ln \alpha_i$。于是，β 直接估计为专利研发弹性，这是因为式(23.31)蕴含 $\partial \ln \mathrm{E}[y_{it} | x_{it}] / \partial \ln x_{it} = \beta$。与对数—对数模型不同，$y_{it}$ 的零值并不会引起任何问题。

更为丰富的参数模型认为，因变量是一种计数形式。其起点是泊松模型：

$$y_{it} | x_{it},\ \gamma_i \sim \mathcal{P}[\exp(\gamma_i + \beta \ln x_{it})] \tag{23.32}$$

该模型将在 23.7 节详述，它与式(23.31)给出的条件均值一样。

表 23.1 列出这些数据的一系列估计量。所有估计量在下面的假设条件下都是一致的，该假设是，条件均值由式(23.31)给出，α_i 随机效应与 $\mathbf{x}_{it}$ 独立，且具有常值均值。除最后一个估计量之外，其他所有估计量在 α_i 固定效应与 $\mathbf{x}_{it}$ 相关的假设条件下均是非一致的。此表提供了三种标准误差估计值：程序默认估计值、面板稳健估计值(若有的话)以及自助法(非精炼的)。每列详细内容如下：

表 23.1　专利与研发支出：非线性面板模型估计量[a]

	NLS	泊松	GEE	泊松-RE	泊松-FE
$\gamma = \ln \alpha$	2.529	1.712	2.068	2.313	—
β	0.509	0.693	0.560	0.349	−0.038
面板 se	(0.055)	(0.043)	(0.033)	(0.033)	(0.033)
方根 se	[0.054]	[0.047]	[0.107]	[0.119]	[0.107]
通常 se	{0.011}	{0.002}	{0.004}	{0.033}	{0.033}
β 求和	—	0.486	0.460	0.546	0.313
N	1 730	1 730	1 730	1 730	1 620

[a] 列出了 ln(专利)对 ln(研发支出)的非线性面板(23.31)混合 NLS、混合泊松、混合 GEE、混合随机效应(RE)、泊松固定效应估计值。斜率系数的标准误差是面板稳健的，由圆括号给出，斜率系数的自助法标准误差由方括号给出，而假设 iid 误差的通常估计的标准误差则由大括号给出。倒数第二行列出含有至多ln(研发支出)5 个滞后期作为回归元的扩展模型中的 β 系数之和。

混合 NLS： 第一列 NLS 估计值是通过 NLS(参见 5.8 节)对满足 $\alpha_i = \alpha$ 的式(23.31) 估计出的。一旦假定 iid 误差，默认标准误差为 0.011，该值远小于正确面板稳健标准误差估计值 0.054。

混合泊松模型： 第二列泊松估计值是通过 MLE 对满足 $\alpha_i = \alpha$ 的泊松模型(23.32)估计出的，这里假定对不同 i 与 t 具有独立性。其估计弹性为 0.693，与之相比，NLS 的估计弹性为 0.509。默认标准误差是 0.002，该值利用了泊松方差均值相等的约束(参见 20.2.2 节)。利用三明治方差矩阵估计(参见 20.2.2 节)对过度分散修正使标准误差估计增大到 0.020，从而使控制计数数据任何过度分散显得具有重要意义。另外，控制对于给定 i 时不同 t 的相关，会导致甚至更高的面板稳健标准误差估计值 0.043。

混合 GEE： 混合 GEE 估计量是式(23.30)的解，其中，$g(\mathbf{x}_{it}, \boldsymbol{\beta})$ 由满足 $\alpha_i = \alpha$ 的式(23.32)给出。这里，对所用的实用矩阵 Σ_i 的特别设定由式(23.55)后面给出。运用后面讨论的面板稳健估计，得到估计弹性为 0.560，其标准误差为 0.033。

泊松 RE： 泊松随机效应估计量，假定 $\alpha_i = \ln \gamma_i$ 服从伽玛分布(参见 23.7.2 节)。估计弹性为 0.349，其默认标准误差为 0.033。

泊松 FE：泊松固定效应估计量，假定 $\alpha_i = \ln \gamma_i$ 是一个固定效应，对它的估计像 23.7.3 节一样。估计弹性为－0.038，它是一个负值，其默认标准误差为 0.033。对于泊松固定效应模型，满足 $\sum_t y_{it}$ 的厂商被省略掉，故使得此模型损失 $22 \times 5 = 110$ 个观测值。

固定效应模型与随机效应模型的结果之间存在很大差异，这支持了固定效应估计。令人惊讶的是，泊松固定效应模型出现了负的估计弹性，原因在于该模型太简单。尤其是，研发支出影响到后期专利活动。当用 $\sum_{l=0}^{5} \beta_l \ln \mathbf{x}_{i,t-1}$ 代替式(23.31)与式(23.32)中的 $\beta \ln \mathbf{x}_{it}$ 时，得出表 23.1 中倒数第二行给出的估计弹性 $\sum_{l=0}^{5} \hat{\beta}_l$。固定效应估计值 0.313 小于其他模型估计值，只是现在差异减小了。

23.4 二值结果数据

我们考察 y_{it} 只取值 0 与 1 的二值结果。例如，数据是某个个体在几个时期的每一个时期是否被雇用。一个重要结果是，固定效应估计对 logit 模型是可解的，却不适合于 probit 模型。

23.4.1 *特定体体效应的二值模型*

具有特定个体效应的二值结果模型的一个正常推广是从横截面数据到面板数据，该模型设定如下：y_{it} 仅取值 0 与 1，满足：

$$\Pr[y_{it} = 1 \mid \mathbf{x}_{it}, \boldsymbol{\beta}, \alpha_i] = \begin{cases} F(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}), & \text{一般形式} \\ \Lambda(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}), & \text{对于 logit 模型} \\ \Phi(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}), & \text{对于 probit 模型} \end{cases} \tag{23.33}$$

其中 $F(\cdot)$ 表示累积分布函数，$\Lambda(\cdot)$ 表示逻辑斯蒂 cdf，满足 $\Lambda(z) = e^z/(1+e^z)$，而 $\Phi(\cdot)$ 表示标准正态累积分布函数。已知式(23.33)，并假定具有条件独立性，第 i 个观测值 $\mathbf{y}_i = (y_{i1}, \cdots, y_{iT})$ 的联合密度是：

$$f(\mathbf{y}_i \mid \mathbf{X}_i, \alpha_i, \boldsymbol{\beta}) = \prod_{t=1}^{T} F(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta})^{y_{it}} (1 - F(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}))^{1-y_{it}} \tag{23.34}$$

对于二值数据，其条件概率也是条件均值，因此：

$$\mathrm{E}[y_{it} \mid \alpha_i, \mathbf{x}_{it}] = F(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} \mid \alpha_i, \mathbf{x}_{it}) \tag{23.35}$$

这是一个单指标特定个体效应模型[参见式(23.5)]，它不可以简化成加法效应模型或乘法效应模型。加法与乘法效应模型都不适合，因为这两种模型都没有将条件均值与条件概率限制于 0 与 1 之间。

由于二值数据必服从贝努利分布，故二值面板模型着重于参数模型(23.34)。条件均值模型(23.35)却极少运用，尽管当回归元为内生的时候使用模型(23.35)会很自然。

23.4.2 *随机效应二值模型*

随机效应极大似然估计量假定，个体效应服从正态分布，$\alpha_i \sim \mathcal{N}[0, \sigma_\alpha^2]$。$\boldsymbol{\beta}$ 与

σ_α^2 的随机效应极大似然估计量是对对数似然函数 $\sum_{i=1}^N \ln f(\mathbf{y}_i \mid \mathbf{X}_i, \boldsymbol{\beta}, \sigma_\alpha^2)$ 求极大值，其中：

$$f(\mathbf{y}_i \mid \mathbf{X}_i, \boldsymbol{\beta}, \sigma_\alpha^2) = \int f(\mathbf{y}_i \mid \mathbf{X}_i, \alpha_i, \boldsymbol{\beta}) \frac{1}{\sqrt{2\pi\sigma_\alpha^2}} \exp\left(\frac{-\alpha_i}{2\sigma_\alpha^2}\right)^2 d\alpha_i \tag{23.36}$$

这里 $f(\mathbf{y}_i \mid \mathbf{X}_i, \alpha_i, \boldsymbol{\beta})$ 已由式(23.34)给出，对于 logit 模型，$F=\Lambda$；而对于 probit 模型，$F=\Phi$。积分(23.36)不存在闭形式解，故一种标准计算方法是对它运用数值求积法。

若固定效应不存在，则替代随机效应模型的一种方法是混合二值模型，该模型直接设定 $\Pr[\mathbf{y}_{it}=1 \mid \mathbf{x}_{it}] = F(\mathbf{x}_{it}'\boldsymbol{\beta})$。于是，统计推断应建立在面板稳健标准误差(参见 23.2.6 节)的基础上。利用广义矩方法可能获得更有效的估计[参见埃弗里等人(Avery et al., 1983)]，广义估计方程参见梁和基格尔(Liang and Zeger, 1986)。

23.4.3 *固定效应 logit*

对于面板 logit 模型，运用条件极大似然估计可能获得固定效应估计，但对于其他二值面板模型诸如面板 probit 却不可以。

就 logit 模型而言，经过 23.4.6 节给出的一些代数运算后，得到 $\mathbf{y}_i = (y_{i1}, \cdots, y_{iT})$ 的联合密度为：

$$f(\mathbf{y}_i \mid \alpha_i, \mathbf{x}_i, \boldsymbol{\beta}) = \frac{\exp(\alpha_i \sum_t y_{it}) \exp((\sum_t y_{it} \mathbf{x}_{it}') \boldsymbol{\beta})}{\prod_t [1 + \exp(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta})]} \tag{23.37}$$

它依赖于 α_i，我们需要剔除 α_i。对于第 i 个观测值，T 个时期存在 1 的结果为 $\sum_t y_{it}$。定义集合 $\mathbf{B}_c = \{\mathbf{d}_i \mid \sum_t d_{it} = \sum_t y_{it} = c\}$ 是 T 个二值结果之和 $\sum_t y_{it} = c$ 的 0 与 1 的所有可能序列集合。于是，当我们以 $\sum_t y_{it} = c$ 为条件，23.4.6 节将要证明，可以剔除 α_i，从而：

$$f(\mathbf{y}_i \mid \sum_t y_{it} = c, \mathbf{x}_i, \boldsymbol{\beta}) = \frac{\exp((\sum_t y_{it} \mathbf{x}_{it}')\boldsymbol{\beta})}{\sum_{\mathbf{d} \in \mathbf{B}_c} \exp((\sum_t d_{it} \mathbf{x}_{it}')\boldsymbol{\beta})} \tag{23.38}$$

该结果归功于张伯伦(Chamberlain, 1980)。密度(23.38)是条件极大似然估计的基础。其唯一复杂的情况是，存在许多集合 $\mathbf{B}_c$ 以及 $\mathbf{B}_c$ 集合之内的序列，正如我们现在所要阐述的。

第一，条件 $\sum_t y_{it} = 0$ 无意义，因为只有所有 $y_{it} = 0$ 时才会如此，类似地，得到 $\sum_t y_{it} = T$ 的情况。这意味着，当大部分人在所有时期都被雇用时，观测值损失相当大。

举以工作为条件的一个例子，假定 $T=2$ 且 $\sum_t y_{it} = 1$。于是，可能情形是序列 $\{0,1\}$ 或是 $\{1,0\}$，例如，由式(23.38)中的条件概率得出：

$$\begin{aligned}\Pr[y_{i1}=0, y_{i2}=1 \mid y_{i1}+y_{i2}=1] &= \frac{\exp(\mathbf{x}_{i1}'\boldsymbol{\beta})}{\exp(\mathbf{x}_{i1}'\boldsymbol{\beta}) + \exp(\mathbf{x}_{i2}'\boldsymbol{\beta})} \\ &= \frac{\exp((\mathbf{x}_{i1} - \mathbf{x}_{i0})'\boldsymbol{\beta})}{1 + \exp((\mathbf{x}_{i1} - \mathbf{x}_{i0})'\boldsymbol{\beta})}\end{aligned}$$

当 $T=3$ 时，我们能以 $\sum_t y_{it} = 1$ 为条件，其可能序列是 $\{0,0,1\}$、$\{0,1,0\}$ 以及 $\{1,0,0\}$，

或者以$\sum_t y_{it}=2$为条件,得出可能序列{0,1,1}、{1,0,1}以及{1,1,0}。很明显,对大T来说,存在众多序列,而且条件密度变得复杂。

条件密度是一种条件logit模型形式,其中,参数是不变的,但回归元对不同选项来说却是变化的。选项数目对不同个体者来说却是变化的,就第i个个体而言,每个选项是0与1的一个特定序列,其序列之和为$\sum_t y_{it}$。利用对此问题的结构进行特定编程是最容易的。甚至选项数目很大时,比如,当$T=10$且$\sum_t y_{it}=5$,就会存在252个选项。通过忽略掉一些观测值,诸如具有许多选项的个体,这是因为$\sum_t y_{it}$很大,或者通过减少时期数目,得到一致但稍欠有效的估计是可能的。

运用最初模型(23.37),若去掉个体效应α_i,会使解释回归系数变得不可能。不过,我们要使用条件模型(23.38)。例如,假定我们拥有单个回归元且$\beta=0.2$。于是,当考察两个时期,并且以$\sum_t y_{it}=1$为条件时,有:

$$\Pr[y_{i1}=0,y_{i2}=1\,|\,y_{i1}+y_{i2}=1]=\frac{\exp(\beta(x_{i1}-x_{i0}))}{1+\exp(\beta(x_{i1}-x_{i0}))}$$

由此可得,x_{i1}与x_{i2}的一个单位差分导致该序列的条件概率变为$\exp(\beta)/[1+\exp(\beta)]$,该值可与当$x_{i1}=x_{i2}$时概率的一半形成对比。

23.4.4 *动态二值模型*

假定我们拥有一个纯时间序列的一阶马尔可夫logit模型,该模型除以下滞后因变量之外没有其他回归元:

$$\Pr[y_{it}=1\,|\,\alpha_i,y_{it-1}]=\frac{\exp(\alpha_i+\gamma y_{it-1})}{1+\exp(\alpha_i+\gamma y_{it-1})} \tag{23.39}$$

执行23.4.6节给出的某些代数运算,得出:

$$f(\mathbf{y}_{it}\,|\,y_{i1},y_{iT},\sum_{t=2}^{T-1}y_{it},\gamma)=\frac{\exp(\gamma\sum_{t=2}^{T-1}y_{it}y_{it-1})}{\sum_{\mathbf{d}\in\mathbf{C}_i}\exp(\gamma\sum_{t=2}^{T-1}d_{it}d_{it-1})} \tag{23.40}$$

其中,集合$\mathbf{C}_i=\{\mathbf{d}_i\,|\,y_{i1},y_{iT},\sum_t d_{it}=\sum_t y_{it}\}$是0与1序列的所有可能集合,这里,$T$个二值结果之和为$\sum_t y_{it}$,其第一个结果是$y_{i1}$,最后的结果则为$y_{iT}$。

建立在式(23.40)基础上的条件ML估计会得出γ的一致估计。要求时期的最小数是4。例如,若$\mathbf{y}_i$是序列{0,1,0,1},则$\mathbf{C}_i$集合由序列{0,1,0,1}与{0,0,1,1}构成。该方法归功于张伯伦(Chamberlain, 1985),他实际上考虑了二阶马尔可夫模型。斋、霍因斯和希斯洛普(Chay, Hoynes and Hyslop, 2001)将此方法应用于加利福尼亚福利时期的管理数据,可以发现,一旦控制不可观测个体异质性,福利分享仍有真实状态相依性。

上述结果与讨论用于纯时间序列模型。奥诺雷和基里亚齐杜(Honoré and Kyriazidou, 2000)给出一种允许除回归元之外没有滞后因变量的方法。因而,假设式(23.39)成为:

$$\Pr[y_{it}=1\,|\,\alpha_i,y_{it-1},\mathbf{x}_{it}]=\frac{\exp(\alpha_i+\mathbf{x}_{it}'\boldsymbol{\beta}+\gamma y_{it-1})}{1+\exp(\alpha_i+\mathbf{x}_{it}'\boldsymbol{\beta}+\gamma y_{it-1})} \tag{23.41}$$

考察四个时期,同时在第1个时期与第4个时期,比如说d_1与d_4,具有共同二值结

果的一种序列。于是，已知该序列要么为$\{d_1,0,1,d_4\}$要么为$\{d_2,1,0,d_4\}$时，该序列是$\{d_1,0,1,d_4\}$ 的概率现在依赖于 α_i。可是，当 $x_{3i}=x_{4i}$时，相关性 α_i 便消失了。由于仅有少数观测值有 $x_{3i}=x_{4i}$，特别是拥有连续数据时，奥诺雷和基里亚齐杜(Honoré and Kyriazidou, 2000)提出了含有依赖于$(x_{3i}-x_{4i})$的核权数的核光滑方法。斋和霍因斯(Çhay and Hyslop, 2000)给出了这种方法的一个应用，以及关于动态二值数据模型的许多其他方法。

23.4.5 *多项式模型*

固定效应估计量能够被推广到多项式 logit 模型，这是因为该模型会得出两两比较选项的二值 logit 模型(参见 15.4.3 节)。对于静态模型，张伯伦(Chamberlain, 1980)给出一个简要解释，而李明宰(M.-J. Lee, 2002)则提出更详细说明。马尼亚克(Magnac, 2000)运用动态固定效应 logit 模型，该模型除滞后因变量以外没有其他回归元，对法国劳动力市场 6 个不同状态之间的个体过渡做出一个相当详细的实证应用。奥诺雷和基里亚齐杜(Honoré and Kyriazidou, 2000)则考察了多项式 logit 模型。

对于其他多项式模型，必须运用随机效应方法。甚至在横截面情况下，对诸如混合 logit 与多项式 probit 这些模型进行估计很复杂。详细内容，参见特雷恩(Train, 2003)。

23.4.6 *固定效应推导*

为了简单起见，不用下标 i。对于 logit 模型，由式(23.34)给出的 $\mathbf{y}=(y_1,\cdots,y_T)$的联合概率变成：

$$\begin{aligned}f(\mathbf{y}\mid\alpha)&=\prod_{t=1}^{T}\left(\frac{\exp(\alpha+\mathbf{x}_t'\boldsymbol{\beta})}{1+\exp(\alpha+\mathbf{x}_t'\boldsymbol{\beta})}\right)^{y_t}\left(\frac{1}{1+\exp(\alpha+\mathbf{x}_t'\boldsymbol{\beta})}\right)^{1-y_t} \qquad (23.42)\\&=\frac{\exp(\sum_t y_t(\alpha+\mathbf{x}_t'\boldsymbol{\beta}))}{\prod_t[1+\exp(\alpha+\mathbf{x}_t'\boldsymbol{\beta})]}\\&=\frac{\exp(\alpha\sum_t y_t)\exp((\sum_t y_t\mathbf{x}_t')\boldsymbol{\beta})}{\prod_t[1+\exp(\alpha+\mathbf{x}_t'\boldsymbol{\beta})]}\end{aligned}$$

由此得到式(23.37)。

可以证明，数量$\sum_t y_t$ 是如下 α 的充分统计量。假定我们拥有 $\mathbf{y}$ 的观测值，使得$\sum_t y_t=c$。定义集合 $\mathbf{B}_c=\{\mathbf{d}\mid\sum_t d_t=c\}$是下述 0 与 1 序列的所有可能情况的集合，该 0 与 1 序列的 T 个二值结果之和为 c，且以$\sum_t y_t=c$ 为条件。从而：

$$\begin{aligned}f(\mathbf{y}\mid\sum_t y_t=c)&=\frac{\Pr[\mathbf{y},\sum_t y_t=c]}{\Pr[\sum_t y_t=c]} \qquad (23.43)\\&=\frac{\Pr[\mathbf{y}]}{\Pr[\sum_t y_t=c]}\\&=\frac{\Pr[\mathbf{y},\sum_t y_t=c]}{\sum_{\mathbf{d}\in\mathbf{B}_c}\Pr[\mathbf{d}]}\\&=\frac{\exp((\sum_t y_t\mathbf{x}_t')\boldsymbol{\beta})}{\sum_{\mathbf{d}\in\mathbf{B}_c}\exp((\sum_t d_t\mathbf{x}_t')\boldsymbol{\beta})}\end{aligned}$$

其中,第一个等式运用了贝叶斯规则。第二个等式用到如下事实:$\Pr[\sum_t y_t = c]$等于 0 与 1 组合为 c 的概率之和。第三个等式用到前面 $f(\mathbf{y})$的定义,以及当我们关注 $\mathbf{d} \in \mathbf{B}_c$ 时,部分地依赖于$\sum_t y_t = \sum_t d_t$ 而产生的大量简化。

现在,考察动态模型。用 γy_{t-1}代替式(23.42)中的 $\mathbf{x}_t'\boldsymbol{\beta}$,得到:

$$f(\mathbf{y}) = \frac{\exp(\alpha \sum_{t=2}^T y_t)\exp(\sum_{t=2}^T \gamma y_{t-1} y_t)}{\prod_t [1+\exp(\alpha+\gamma y_{t-1})]}$$

$$= \frac{\exp(\alpha \sum_{t=2}^T y_t)\exp(\sum_{t=2}^T \gamma y_{t-1} y_t)}{[1+\exp(\alpha)]^{\sum_{t=2}^T 1-y_{t-1}}[1+\exp(\alpha+\gamma)]^{\sum_{t=2}^T y_{t-1}}}$$

$$= \frac{\exp(\alpha \sum_{t=2}^T y_t)\exp(\sum_{t=2}^T \gamma y_{t-1} y_t)}{[1+\exp(\alpha)]^{(T-1+y_1-y_T+\sum_{t=2}^T y_t)}[1+\exp(\alpha+\gamma)]^{y_1-y_T+\sum_{t=2}^T y_t}}$$

其中,第二个等式用到了 y_{t-1}为 1 或为 0 的事实,随后进行某些代数运算,最后等式则用到$\sum_{t=2}^T y_{t-1} = y_1 - y_T + \sum_{t=2}^T y_t$。然后,类似于式(23.43)那样,经过一些代数运算,除了$\sum_{t=2}^T y_t$ 的条件之外,还需要以分母出现 y_1 与 y_T 为条件。等价地讲,我们可以把$\sum_{t=1}^T y_t$ 与 y_1 及 y_T 为条件。从而得出:

$$f(\mathbf{y}) = \frac{\exp(\sum_{t=2}^T \gamma y_{t-1} y_t)}{\sum_{\mathbf{d}\in \mathbf{C}_c} \exp(\sum_{t=2}^T \gamma d_{t-1} d_t)}$$

其中,$\mathbf{C} = \{\mathbf{d} | d_1 = y_1, d_T = y_T, \sum_{t=1}^T d_t = \sum_{t=1}^T y_t\}$是下述 0 与 1 序列的所有可能情况集合,这 T 个 0 与 1 序列的二值结果之和是 $\sum_t y_t$,其第一个结果是 y_1,而最后一个结果是 y_T。

23.5 Tobit 模型与选择模型

当可以利用面板数据而不是单一横截面数据时,考察删失、截尾或者选择模型。

混合分析直接反映出横截面情况下的分析,对面板稳健标准误差的计算应该加以调整(参见 23.2.8 节)。例如,参见格拉斯德尔(Grasdal,2001),他考虑了由面板损耗引起的选择。

不过,这里我们关注含有特定个体效应的面板模型。若能保证对纯随机效应做出强假设,就能估计随机效应模型,其唯一困难就是进行数值计算。不过,在短面板的通常微观经济计量学背景下,固定效应模型不存在简单的一致估计量。对于 23.8 节给出的 Tobit 模型与广义 Tobit 模型中的固定效应来说,可能得出更为复杂的半参数估计量。

23.5.1 *截删与截尾模型*

对于横截面数据,删失 23.5.1 模型已由 16.3.1 节给出。一种具有可加特定个体效应的面板形式设定如下:

$$y_{it}^* = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} \tag{23.44}$$

其中，$\varepsilon_{it} \sim \mathcal{N}[0,\sigma_\varepsilon^2]$，当 $y_{it}^*>0$ 或 $y_{it}^*=0$ 时，我们观测到 $y_{it}=y_{it}^*$，而当 $y_{it}^*\leqslant 0$ 时，则不可观测。第 i 个观测值的联合密度能够被写成：

$$f(\mathbf{y}_i \mid \mathbf{X}_i,\alpha_i,\boldsymbol{\beta},\sigma_\varepsilon^2) = \prod_{t=1}^{T}\left[\frac{1}{\sigma_\varepsilon}\phi_{it}\right]^{d_{it}}[1-\phi_{it}]^{1-d_{it}} \tag{23.45}$$

其中，$\phi_{it}=\phi(y_{it}-\alpha_i-\mathbf{x}_{it}'\boldsymbol{\beta}/\sigma_\varepsilon)$，$\Phi_{it}=\Phi((\alpha_i+\mathbf{x}_{it}'\boldsymbol{\beta})/\sigma_\varepsilon)$，而 $\phi(\cdot)$ 与 $\Phi(\cdot)$ 分别表示标准正态的 pdf 与 cdf。

固定效应极大似然估计是求建立在式(23.45)基础上的对数似然关于 $\boldsymbol{\beta},\sigma_\varepsilon^2,\alpha_1,\cdots,\alpha_N$ 的极大值。在短面板情况下，所得到的 $\boldsymbol{\beta}$ 一致估计量是非一致的，原因在于，存在非主要参数问题，而且没有简单的差分方法或条件方法能够提供一致估计量。赫克曼和麦柯迪(Heckman and MaCurdy, 1980)将固定效应 MLE 用于妇女劳动供给。尽管认识到估计量有非一致性，他们仍讨论了当 $T=8$ 时非一致性可能不是太大。格林(Greene, 2004a)对固定效应 Tobit MLE 做出一个最新的蒙特卡罗研究。

由于固定效应估计量存在非一致性，故人们更广泛地运用随机效应估计。在 $\alpha_i \sim \mathcal{N}[0,\sigma_\alpha^2]$ 假设下，$\boldsymbol{\beta}$、σ_ε^2 以及 σ_α^2 的随机效应 MLE 是对对数似然 $\sum_{i=1}^N \ln f(\mathbf{y}_i \mid \mathbf{X}_i,\boldsymbol{\beta},\sigma_\varepsilon^2,\sigma_\alpha^2)$ 求极大值。

$$f(\mathbf{y}_i \mid \mathbf{X}_i,\boldsymbol{\beta},\sigma_\varepsilon^2,\sigma_\alpha^2) = \int f(\mathbf{y}_i \mid \mathbf{X}_i,\alpha_i,\boldsymbol{\beta},\sigma_\varepsilon^2)\frac{1}{\sqrt{2\pi\sigma_\alpha^2}}\exp\left(\frac{-\alpha_i}{2\sigma_\alpha^2}\right)^2 d\alpha_i \tag{23.46}$$

其中，$f(\mathbf{y}_i \mid \mathbf{X}_i,\alpha_i,\boldsymbol{\beta},\sigma_\varepsilon^2)$ 已由式(23.45)给出。利用高斯求积法能计算该一维积分。

此方法能推广到含有删失或截尾的其他模型上。例如，当 10 以上计数仅被记录成 10 或稍大一些的数时，就可运用 23.7.2 节中泊松随机效应模型的右删失形式。

完全参数方法有两个弱点。第一，如同横截面情况一样，当存在删失或截尾时，依赖于分布假设的程度会更大。第二，关于纯随机效应的回归元是独立的假设可能显得太强。

23.5.2 *选择模型*

由于面板数据具有类似于横截面数据导致选择模型的原因(参见 16.5 节)，所以面板数据同样会出现选择模型问题。16.5.1 节的第二类 Tobit 模型被推广到含有特定个体效应 λ_i 与 δ_i 的线性面板模型是：

$$\begin{aligned} y_{it}^* &= \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} \\ d_{it}^* &= \delta_i + \mathbf{z}_{it}'\boldsymbol{\gamma} + v_{it} \end{aligned} \tag{23.47}$$

其中，当 $d_{it}^*>0$ 时，$y_{it}=y_{it}^*$ 是可观测的。否则，y_{it} 是不可观测的。

对于随机效应系统表述来说，假定四种不可观测因素服从正态分布。豪斯曼和怀斯(Hausman and Wise, 1979)提出一种极大似然估计，由于 α_i 与 δ_i 可能相关，并且 ε_{it} 可能与 v_{it} 相关，故这是二变量积分。

在短面板条件下，固定效应估计量是非一致的。不过，注意到，当 $d_{it}^*=\delta_i$ 归因

于个体时常值特性引起选择可能是可观测的或不可观测的，则模型 $y_{it}=\alpha_i+\mathbf{x}_{it}'\boldsymbol{\beta}+\varepsilon_{it}$ 的固定效应估计量是一致的。固定效应面板模型控制住了样本选择，目的在于它依赖时常值特性。

费尔贝克和尼吉曼（Verbeek and Nijman，1992）对这些模型进行一致估计所需要的基本假设做出更详细讨论，并提出有关选择偏倚的检验。伍德里奇（Wooldridge，1995）在较弱假设下做出了类似分析，并阐述在允许固定效应模型有一致估计的一些应用中可能不是限制太强的假设。维拉（Vella，1998）提供了一个评注，以及其他一些参考文献。

样本选择方法被推广到面板损耗（参见 21.8.5 节），从而导致当因变量观测值以非随机方式丢失时出现损耗偏倚。于是，当 $d_{it}^*\leqslant 0$ 时，第 i 个观测值的所有数据都不是可观测到的，所以式（23.47）中的 $\mathbf{z}_{it}$ 需要用不同于时期 t 的一些时期中的观测变量来代替。一个早期例子是由豪斯曼与怀斯（Hausman and Wise，1979）给出的，而最近应用则由格拉斯德尔（Grasdal，2001）提出。巴尔塔基（Baltagi，2001）与萧政（Hsiao，2003）给出了更多的参考文献。

23.6 过渡数据

为了具体起见，考察有关福利时期的面板数据。最大关注点是测算福利时期方面的个体持久性，并确定个体持久性归因于真实状态相依性的范围，而不是因福利引起的个体习性差异。由于个体习性可能部分地依赖于不可观测因素，故应该使用具有特定个体效应的模型。对于持续期限数据，建模方法异常丰富，原因在于可以利用的过渡面板数据可能有多种类型。此处，我们关注固定效应模型。

可以利用有关某个个体是否处于一个状态的几个时点上诸如福利情况的数据。那么，人们就能使用二值面板模型（参见 23.4 节），比如动态固定效应 logit 模型。

较丰富数据提供了几个个体时期的持续期限方面的信息。一个通常起点是，面板比例风险模型：

$$\lambda(t_{ij}\,|\,\mathbf{x}_{ij})=\lambda_j(t_{ij},\gamma_j)\exp(\mathbf{x}_{ij}'\boldsymbol{\beta})\alpha_i \tag{23.48}$$

其中，t_{ij} 表示第 i 个个体的第 j 个时期的完整时期持续期限，α_i 表示特定个体效应。这是一个混合比例风险模型，第 18 章曾经讨论过关于单时期数据的此类模型。关于仅有单时期数据的 MPH 模型的非参数识别的条件，包括 α_i 作为回归元的独立分布的假设。这就剔除了固定效应。不过，若有多重时期可利用，奥诺雷（Honoré，1992）已经证明，当 $\mathbf{x}_{ij}$ 对不同 j 而言都是常值时，α_i 可以是固定效应（参见 19.4.1 节）。有关模型（23.48）的进一步讨论，包括对依赖于第一个时期持续期限的第二个时期来说含有风险函数的动态持续期限模型，参见 19.4.1 节。

张伯伦（Chamberlain，1985）阐述了各种面板持续期限模型中剔除 α_i 的几种方法。对于 MPH 模型，其基线风险 $\lambda_j(\cdot)$ 在不同时期 j 都相同时，第二个时期比第一个时期更长的概率并不依赖于 α_i。条件极大似然法能应用于伽玛持续期限，

因为伽玛是 LEF 密度。对于威布尔、伽玛以及对数正态模型，t_{i1}/t_{i2} 的密度不依赖于 α_i。

关于最新参考文献以及详细讨论，包括多重时期数据对删失的敏感性，可参见范登堡(Van den Berg, 2001)。

23.7 计数数据

豪斯曼等人(Hausman et al., 1984)已经阐述可估计的固定效应模型与随机效应模型，这里的模型既有面板泊松模型，又有面板负二项模型。最新的研究工作强调乘法效应模型的固定效应，在相对弱分布的假设下，以便得出静态模型及动态模型的估计。

23.7.1 *特定个体效应的计数模型*

尽管负二项式模型的面板形式已经被简要地讨论，20.2 节详细分析横截面数据的情况，但我们仍要关注泊松模型。

泊松特定个体效应模型就是设定 $y_{it} \sim \mathcal{P}[\alpha_i \exp(\mathbf{x}_{it}'\boldsymbol{\beta})]$。于是，若假定具有条件独立性，则第 i 个观测值 $\mathbf{y}_i=(y_{i1},\cdots,y_{iT})$ 的联合密度是：

$$f(\mathbf{y}_i \mid \mathbf{X}_i,\alpha_i,\boldsymbol{\beta})=\prod_{t=1}^{T}\exp[-\alpha_i\exp(\mathbf{x}_{it}'\boldsymbol{\beta})][\alpha_i\exp(\mathbf{x}_{it}'\boldsymbol{\beta})]^{y_{it}}/y_{it}! \tag{23.49}$$

一种较少参数的方法直接将条件均值建模成：

$$\begin{aligned}\mathrm{E}[y_{it}\mid\alpha_i,\mathbf{x}_{it}]&=\alpha_i\exp(\mathbf{x}_{it}'\boldsymbol{\beta})\\&=\exp(\gamma_i+\mathbf{x}_{it}'\boldsymbol{\beta})\end{aligned} \tag{23.50}$$

这既是一个单指标特定个体效应模型，又是一个乘法效应模型。由于该模型是一个乘法效应模型，故个体效应 α_i 可通过均值差分或一阶差分加以剔除。注意，泊松面板模型(23.49)具有条件均值(23.50)。

23.7.2 *随机效应计数模型*

假定服从伽玛分布的随机效应会得出易于处理的随机效应模型的边际密度。假如 α_i 服从 $\mathcal{G}[\eta,\eta]$ 分布，其均值为 1，方差为 $1/\eta$，密度 $g(\alpha_i\mid\eta)=\eta^{\eta}\alpha_i^{\eta-1}e^{-\alpha_i\eta}/\Gamma(\eta)$。从而，对于泊松模型(23.49)来说，式(23.18)变成：

$$f(\mathbf{y}_i\mid\mathbf{X}_i,\boldsymbol{\beta},\eta)=\left[\prod_t\frac{\lambda_{it}^{y_{it}}}{y_{it}!}\right]\times\left(\frac{\eta}{\sum_t\lambda_{it}+\eta}\right)^{\eta}\times\left(\sum_t\lambda_{it}\right)^{-\sum_t y_{it}}\frac{\Gamma(\sum_t y_{it}+\eta)}{\Gamma(\eta)} \tag{23.51}$$

其中，$\lambda_{it}=\exp(\mathbf{x}_{it}'\boldsymbol{\beta})$，23.7.5 节将给出其推导。所得到的泊松随机效应估计量 $\hat{\boldsymbol{\beta}}$ 的一阶条件能表述成：

$$\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{x}_{it}\left(y_{it}-\lambda_{it}\frac{\bar{y}_i+\eta/T}{\bar{\lambda}_i+\eta/T}\right)=\mathbf{0} \tag{23.52}$$

其中，$\bar{\lambda}_i = T^{-1}\sum_t \exp(\mathbf{x}_{it}'\boldsymbol{\beta})$。

当以所有时期回归元为条件的均值 $\mathrm{E}[y_{it}|\alpha_i,\mathbf{x}_{i1},\cdots,\mathbf{x}_{iT}]=\alpha_i\exp(\mathbf{x}_{it}'\boldsymbol{\beta})$时，式(23.52)左边项具有期望值 0。因此，虽然做出全部参数假设，但在相对弱假设——条件均值是式(23.50)给出的形式且回归元是强外生的——条件下，泊松随机效应估计量关于 $\boldsymbol{\beta}$ 是一致的。对于密度(23.51)，$\mathrm{E}[y_{it}|\mathbf{x}_i]=\lambda_{it}$，而 $\mathrm{V}[y_{it}|\mathbf{x}_i]=\lambda_{it}+\lambda_{it}^2/\delta$，所以 NB2 形式为过度分散的。其方差矩阵的三明治估计使得对过度分散与条件相关实施更灵活建模成为可行的。尽管信息矩阵关于 $\boldsymbol{\beta}$ 与 η 是分块对角的，但 η 的一阶条件(没有给出)则是相当复杂的。

已知随机效应，有几种可供选择的估计量可以利用。第一，混合泊松估计量忽略了随机效应，并假定 $y_{it}|\mathbf{x}_{it}\sim\mathcal{P}[\exp(\mathbf{x}_{it}'\boldsymbol{\beta})]$。这就得出一阶条件：

$$\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{x}_{it}(y_{it}-\lambda_{it})=\mathbf{0} \tag{23.53}$$

其中，$\lambda_{it}=\exp(\mathbf{x}_{it}'\boldsymbol{\beta})$。若其条件均值为式(23.50)，满足 $\mathrm{E}[\alpha_i|\mathbf{x}_{it}]=1$，则该估计量是一致的。因而，当真实模型是一个具有乘法随机效应的模型时，通常横截面泊松极大似然估计量就是一致的。不过，正如 23.3 节例子所阐明的，应该使用面板稳健标准误差。从而，由式(23.26)得出：

$$\hat{\mathrm{V}}[\hat{\boldsymbol{\beta}}_{\text{混合}}]=\left[\sum_{i,t}\hat{\lambda}_{it}\mathbf{x}_{it}\mathbf{x}_{it}'\right]^{-1}\sum_{i,t,s}\hat{u}_{it}\hat{u}_{is}\mathbf{x}_{it}\mathbf{x}_{it}'\left[\sum_{i,t}\hat{\lambda}_{it}\mathbf{x}_{it}\mathbf{x}_{it}'\right]^{-1} \tag{23.54}$$

其中，$\hat{\lambda}_{it}=\exp(\mathbf{x}_{it}'\boldsymbol{\beta})$，$\hat{u}_{it}=y_{it}-\hat{\lambda}_{it}$，$\sum_{i,t}$ 表示 $\sum_{i=1}^{N}\sum_{t=1}^{T}$，而 $\sum_{i,t,s}$ 表示 $\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}$。一种可供选择的建立在式(23.50)基础上的混合估计量，是非线性最小二乘法(NLS)，在这种情况下，式(23.53)变成$\sum_i\sum_t\mathbf{x}_{it}\lambda_{it}(y_{it}-\lambda_{it})=\mathbf{0}$。

第二，利用 23.2.8 节的广义估计方程方法可能获得更有效的混合估计，该节介绍了条件相关性。对于 $g_{it}=\lambda_{it}=\exp(\mathbf{x}_{it}'\boldsymbol{\beta})$，一般结果(23.30)变成：

$$\sum_{i=1}^{N}\mathbf{Z}_i'\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i-\boldsymbol{\lambda}_i)=\mathbf{0} \tag{23.55}$$

其中，$\mathbf{Z}_i$ 表示 $T\times K$ 阶矩阵，其第 t 行观测值为$\lambda_{it}\mathbf{x}_{it}'$，而 $\boldsymbol{\lambda}_i$表示 $T\times1$ 维向量，其第 t 个元素为λ_{it}。$\mathrm{V}[\mathbf{y}_i|\mathbf{X}_i]$有几种不同的实用方差矩阵 $\boldsymbol{\Sigma}_i$ 可以运用。选取 $\boldsymbol{\Sigma}_i=\mathrm{Diag}[\lambda_{it}]$会得到式(23.53)的混合泊松估计方程。一旦令 $\Sigma_{i,tt}=\lambda_{it}$ 且 $\Sigma_{i,ts}=\lambda_{is}=\phi\sqrt{\lambda_{it}\lambda_{is}}$，对于 $s\neq t$ 时，这允许出现对不同 t 为等相关的或者可交换的相关性，因为该相关行是一个常值 ϕ，对于 $s\neq t$。

第三，利用以负二项式而不是泊松模型作为起点的极大似然法，可以获得更有效的混合估计。假定 y_{it}是一个具有 NB2 方差函数的 iid 负二项式，该 NB2 方差函数的参数为 $\alpha_i\lambda_{it}$ 与 ϕ_i(参见 20.4.1 节)，这蕴含 y_{it}具有均值$\alpha_i\lambda_{it}/\phi_i$，其方差为$(\alpha_i\lambda_{it}/\phi_i)\times(1+\alpha_i/\phi_i)$。若$(1+\alpha_i/\phi_i)^{-1}$是一个参数为$(\eta_1,\eta_2)$且服从贝塔分布的随机变量，则经过一些大量代数运算后，式(23.18)简化成：

$$f(\mathbf{y}_i|\mathbf{X}_i,\boldsymbol{\beta},\eta)=\left(\prod_t\frac{\Gamma(\lambda_{it}+y_{it})!}{\Gamma(\lambda_{it})!\Gamma(y_{it}+1)!}\right)\times\frac{\Gamma(\eta_1+\eta_2)\Gamma(\eta_1+\sum_t\lambda_{it})\Gamma(\eta_2+\sum_t y_{it})}{\Gamma(\eta_1)\Gamma(\eta_2)\Gamma(\eta_1+\eta_2+\sum_t\lambda_{it}+\sum_t y_{it})} \tag{23.56}$$

其中，$\lambda_{it}=\exp(\mathbf{x}_{it}'\boldsymbol{\beta})$。这是对 $\boldsymbol{\beta}$、η_1 以及 η_2 进行极大似然估计的基础。该模型依赖的假设比泊松随机效应模型所需的假设更强。

第四，不必将分析限制在 $f(\mathbf{y}_i|\mathbf{X}_i,\boldsymbol{\beta},\eta)$ 的具有闭形式解的参数模型上。克雷蓬和迪热特（Crépon and Dugeut, 1997a）运用模拟极大似然方法对含有正态随机效应的围栏与零膨胀面板计数模型进行了估计。

23.7.3 固定效应计数模型

泊松面板模型（23.50）的固定效应估计量能以几种不同方式推导出来。

第一种方式为，利用泊松极大似然估计法联立估计法估计 $\boldsymbol{\beta}$ 与 $\alpha_1,\cdots,\alpha_N$。建立在式（23.49）基础上的对数似然是：

$$\begin{aligned}\ln \mathrm{L}(\boldsymbol{\beta},\boldsymbol{\alpha}) &= \ln\Big[\prod_i\prod_t\{\exp(-\alpha_i\lambda_{it})(\alpha_i\lambda_{it})^{y_{it}}/y_{it}!\}\Big] \\ &= \sum_i\Big[-\alpha_i\sum_t\lambda_{it}+\ln\alpha_i\sum_t y_{it}+\sum_t y_{it}\ln\lambda_{it}-\sum_t\ln y_{it}!\Big]\end{aligned} \tag{23.57}$$

其中，$\lambda_{it}=\exp(\mathbf{x}_{it}'\boldsymbol{\beta})$。求上式关于 α_i 的导数，并令该导数为 0，得出 $\hat{\alpha}_i=\sum_t y_{it}/\sum_t\lambda_{it}$。然后，将此 $\hat{\alpha}_i$ 代入式（23.57），就得到中心似然函数（**concentrated likelihood function**）。当省略不包含 $\boldsymbol{\beta}$ 的一些项，则得出：

$$\ln \mathrm{L}_{\mathrm{conc}}(\boldsymbol{\beta})\propto\sum_i\sum_t\Big[y_{it}\ln\lambda_{it}-y_{it}\ln\Big(\sum_s\lambda_{is}\Big)\Big] \tag{23.58}$$

由此可得，对于泊松固定效应模型，不存在非主要参数问题。当固定 T 且 $N\to\infty$ 时，通过对式（23.58）的 $\ln \mathrm{L}_{\mathrm{conc}}(\boldsymbol{\beta})$ 求极大值，获得 $\boldsymbol{\beta}$ 的一致估计值。对式（23.58）求关于 $\boldsymbol{\beta}$ 的导数，得到：

$$\sum_i\sum_t\Big[y_{it}\mathbf{x}_{it}-y_{it}\Big[\sum_s\lambda_{is}\mathbf{x}_{is}\Big]\Big/\Big[\sum_s\lambda_{is}\Big]\Big]=\mathbf{0}$$

然后对上式重新写成：

$$\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{x}_{it}\Big(y_{it}-\frac{\lambda_{it}}{\bar{\lambda}_i}\bar{y}_i\Big)=\mathbf{0} \tag{23.59}$$

其中，$\lambda_{it}=\exp(\mathbf{x}_{it}'\boldsymbol{\beta})$，$\bar{\lambda}_i=T^{-1}\sum_t\exp(\mathbf{x}_{it}'\boldsymbol{\beta})$；参见布伦德尔、格里菲思和温德梅杰（Blundell, Griffith and Windmeijer, 1995）。泊松面板模型（23.49）与 21.6 节的线性面板模型都是少有的，因为对 $\boldsymbol{\beta}$ 与 $\boldsymbol{\alpha}$ 的联立估计在短面板条件下会得出 $\boldsymbol{\beta}$ 的一致估计值，故不存在非主要参数问题。

第二种方式为，条件极大似然估计法通过以 α_i 的充分统计量为条件来剔除固定效应。对于泊松面板模型，充分统计量是 $\sum_t y_{it}$。运用 23.7.5 节给出的一些代数运算可以证明。这就得出与式（23.58）给出的中心对数似然函数成比例的条件对数似然函数。由此可得，固定效应泊松模型 $\boldsymbol{\beta}$ 的条件极大似然估计量是式（23.89）的解。这是由帕姆格伦（Palmgren, 1981）与豪斯曼等人（Hausman et al., 1984）提出的关于 $\boldsymbol{\beta}$ 的泊松固定效应估计量的最初推导。

第三种方式为，对乘法效应模型（23.50）运用均值差分变换（23.14），得出

$E[y_{it}-(\lambda_{it}/\bar{\lambda}_i)\bar{y}_i|\mathbf{x}_{i1},\cdots,\mathbf{x}_{iT}]=0$，从而：

$$E[\mathbf{x}_{it}(y_{it}-(\lambda_{it}/\bar{\lambda}_i)\bar{y}_i)]=\mathbf{0} \tag{23.60}$$

利用相应的样本矩条件，得出作为式(23.59)解的估计量 $\boldsymbol{\beta}$。

同样的估计量可用三种不同方式获得。第三种方式推导很明显做出，使泊松固定效应估计量成为一致的基本假设：回归元是强外生的且式(23.50)被正确设定。推断应建立在面板稳健标准误差的基础上。尤其是，当运用通常默认极大似然或条件极大似然输出时，由前两种推导，由于控制计数数据过度分散失败，所以标准误差可能被大大低估了。固定效应估计量导致了数据的某种损失，因为满足 $\sum_t y_{it}=0$ 的观测值并没有贡献于式(23.59)中的和。

对于负二项式模型的特殊参数化，在有固定效应条件下，还可能获得 $\boldsymbol{\beta}$ 的一致估计。豪斯曼等人(Hausman et al.，1984)曾经假定，y_{it}是 iid 的 NB1，其参数为 $\alpha_i\lambda_{it}$ 与 ϕ_i，这里，$\lambda_{it}=\exp(\mathbf{x}_{it}'\boldsymbol{\beta})$，故 y_{it} 具有均值 $\alpha_i\lambda_{it}/\phi_i$，其方差为 $(\alpha_i\lambda_{it}/\phi_i)\times(1+\alpha_i/\phi_i)$。参数 α_i 与 ϕ_i 仅在至多差一个比值 α_i/ϕ_i 的条件下是可识别的，而且该 α_i/ϕ_i 比值从第 i 个观测值的条件联合密度中消失，经过一些代数运算后，可以证明：

$$f(y_{i1},\cdots,y_{iT}|\sum_t y_{it})=\left(\prod_t\frac{\Gamma(\lambda_{it}+y_{it})}{\Gamma(\lambda_{it})\Gamma(y_{it}+1)}\right)\times\frac{\Gamma(\sum_t\lambda_{it})\Gamma(\sum_t y_{it}+1)}{\Gamma(\sum_t\lambda_{it}+\sum_t y_{it})} \tag{23.61}$$

整数 λ_{it}的这个分布是负二项超几何分布。$\boldsymbol{\beta}$ 的条件极大似然负二项式固定效应估计量是求基于式(23.61)的对数似然函数的极大值。人们更广泛地使用泊松固定效应模型，因为在更弱分布的假设下，$\boldsymbol{\beta}$ 估计量是一致的。

23.7.4 *动态计数模型*

将动态特性引入计数数据模型之中有几种方法。卡梅伦和特里维迪(Cameron and Trivedi，1998)对纯时间序列模型已经给出了一个综述。为了简单起见，考虑包含一阶滞后因变量。一个明显模型是 $E[y_t|y_{t-1},\mathbf{x}_t]=\exp(\gamma y_{t-1}+\mathbf{x}_t'\boldsymbol{\beta})$，可是由于出现 y_{t-1} 的幂而引发了迅速扩大的特性。不过，一个更稳定的模型可通过用 $\exp(\gamma\ln y_{t-1}+\mathbf{x}_t'\boldsymbol{\beta})$来获得，但当 $y_{t-1}=0$ 时，就会遇到问题。因此，一个引人注目的模型是线性反馈模型 $E[y_t|y_{t-1},\mathbf{x}_t]=\gamma y_{t-1}+\exp(\mathbf{x}_t'\boldsymbol{\beta})$。泊松整数值 AR(1)模型具有这种性质，而在纯时间序列情况下，类似于 AR(1)模型，具有相关函数 $\mathrm{Cor}[y_t,y_{t-k}]=\gamma^k$[参见阿洛施和阿尔萨德(Al-Osh and Alzaid，1987)]。

因而，布伦德尔、格里菲思和温德梅杰(Blundell，Griffiths，and Windmeijer，1995，2002)考察了满足

$$E[y_{it}|\alpha_i,y_{i,t-1},\mathbf{x}_{it}]=\gamma y_{i,t-1}+\alpha_i\exp(\mathbf{x}_{it}'\boldsymbol{\beta})$$

的动态固定效应面板数据模型。应用一阶差分变换式(23.15)，得出条件矩约束：

$$E\left[\frac{\exp(\mathbf{x}_{i,t-1}'\boldsymbol{\beta})}{\exp(\mathbf{x}_{it}'\boldsymbol{\beta})}(y_{it}-\gamma y_{i,t-1})-(y_{i,t-1}-\gamma y_{i,t-2})|y_{i1},\cdots,y_{i,t-2},\mathbf{x}_{i1},\cdots,{}_{i,t-1}\right]=0$$

如同 23.2.6 节，该式会产生许多无条件矩条件（参见 22.5.3 节对线性模型的类似讨论），这就为进行广义矩估计提供了基础。克雷蓬和迪热特（Crépon and Dugeut, 1997b），蒙塔尔沃（Montalvo, 1997），布伦德尔、格里菲思和范·里宁（Blundell, Griffith, and Van Reenen, 1999）运用类似的拟差分法，应用于专利与研发关系的研究。

伯肯霍尔特（Böckenholt, 1999）运用更加参数化的模型，使用有限混合分布估计了含有不可观测异质性的泊松整数值 AR(1) 模型（参见 18.5 节）。

23.7.5 随机效应与固定效应泊松模型推导

首先，考察含有随机效应服从伽玛分布的一种随机效应泊松模型。为了简单起见，不用下标 i，并设 $\lambda_t=\exp(\mathbf{x}_t'\boldsymbol{\beta})$。对于泊松模型(23.49)以及随机效应密度 $g(\alpha\mid\gamma)$ 来说，由一般公式(23.18)得出：

$$\begin{aligned}f(y_1,\cdots,y_T\mid\mathbf{x}_t)&=\int_0^\infty\Big[\prod_t(e^{-\alpha\lambda_t}(\alpha\lambda_t)^{y_t}/y_t!)\Big]g(\alpha\mid\gamma)d\alpha\\&=\int_0^\infty\Big[\prod_t\lambda_t^{y_t}/y_t!\Big](e^{-\alpha\sum_t\lambda_t}\cdot\alpha^{\sum_t y_t})g(\alpha\mid\gamma)d\alpha\\&=\Big[\prod_t\lambda_t^{y_t}/y_t!\Big]\times\int_0^\infty(e^{-\alpha\sum_t\lambda_t}\cdot\alpha^{\sum_t y_t})g(\alpha\mid\gamma)d\alpha\end{aligned}$$

对于 $g(\alpha_i\mid\eta)=\eta^\eta\alpha^{\eta-1}e^{-\alpha\eta}/\Gamma(\eta)$，执行类似于 20.4.1 节的那些代数运算，得到由式(23.51)给出的密度。

其次，就已知个体而言，为了简单起见，将个体下标 i 省略，对所有时期观测值来说，推导泊松固定效应模型的条件密度。一般地讲，给定 $\sum_t y_t$ 时 $y_1,\cdots,y_T$ 的密度是：

$$\begin{aligned}f(y_1,\cdots,y_T\mid\sum_t y_t)&=f(y_1,\cdots,y_T\mid\sum_t y_t)/f(\sum_t y_t)\\&=f(y_1,\cdots,y_T)/f(\sum_t y_t)\\&=\frac{\prod_t(\exp(-\mu_t)\mu_t^{y_t}/y_t!)}{\exp(-\sum_t\mu_t)(\sum_t\mu_t)^{\sum_t y_t}/(\sum_t y_t)!}\\&=\frac{\exp(-\sum_t\mu_t)\prod_t\mu_t^{y_t}/\prod_t y_t!}{\exp(-\sum_t\mu_t)\prod_t(\sum_s\mu_s)^{y_t}/(\sum_t y_t)!}\\&=\frac{(\sum_t y_t)!}{\prod_t y_t!}\times\prod_t\Big(\frac{\mu_t}{\sum_s\mu_s}\Big)^{y_t}\end{aligned}$$

其中，第二个等式用到了已知 $y_1,\cdots,y_T$ 的知识，$\sum_t y_t$ 的知识并不会增加什么内容的这个事实，第三个等式是对 y_t iid $\mathcal{P}[\mu_t]$ 加以专门研究，从而 $\sum_t y_t$ 服从 $\mathcal{P}[\sum_t\mu_t]$，而第四个等式与第五个等式则是通过简化得出的。其条件密度 $\sum_t y_t$ 是关于试验的多项式形式，其中，T 个不同结果中的第 t 个结果以概率 $\mu_t/\sum_s\mu_s$ 出现在任何试验中。设 $\mu_{it}=\alpha_i\exp(\mathbf{x}_t'\boldsymbol{\beta})$，并取对数，得出与式(23.58)给出的中心对数似然成比例的条件似然。

23.8 半参数估计

面板数据的半参数文献强调受限因变量的模型，至于横截面数据，当出现截尾、删失或者选择时，参数假设变得尤为重要。关注焦点是含有固定效应的模型。对此，我们给出一个简略概述。

对于二值数据，曼斯基(Manski，1987)将其极大得分估计量从横截面模型推广到具有式(23.33)给出的固定效应的面板模型上，现在函数 $F(\cdot)$不再被设定。尽管该估计量是一致的，但它的收敛速度比$\sqrt{N}$慢，并且不服从渐近正态分布。

对于 Tobit 模型，奥诺雷(Honoré，1992)将鲍威尔(Powell，1986a)的删失 LAD 方法推广到面板效应模型(23.45)，其中，误差项 ε_{it}的分布是未设定的。对数据要加以调整，以便随后通过适当差分剔除固定效应。这种估计量是$\sqrt{N}$一致的，且服从渐近正态分布。

对于含有样本选择的面板数据，基里亚齐杜(Kyriazidou，1997)考察了第二种类型 Tobit 模型的固定效应形式，其中，误差 ε_{it} 与 v_{it} 的分布均未设定。她阐述了赫克曼形式的两步估计量。曼斯基(Manski，1987)极大得分估计量的光滑形式可剔除选择方程中的固定效应，尽管为了剔除结果方程的固定效应，要在第二阶段使用相当复杂的差分方法。该方法能被推广到其他广义 *Tobit* 模型上，沙利耶、梅伦伯格和范泽斯特(Charlier，Melenberg，and van Soest，2001)给出罗伊模型的面板形式或第五种类型 Tobit 模型的一种应用。

持续期限模型普遍都有删失。23.6 节关注完整时期的面板模型。不论是完整时期还是不完整时期就个体而言均可观测时，由于已知存在时不变的固定效应时删失不是独立的，故偏似然方法是不合适的。霍罗威茨和李(Horowitz and Lee，2004)提出含有不完整时期的 MPH 模型(23.43)的一致估计量，该方法并不要求设定基线风险。

23.9 应用研究

正如线性模型情况一样，若使用面板数据，则至少需要推断其建立在面板稳健标准误差基础上。对于横截面数据来说，计算机程序是不会提供这些内容的，除非计算机程序有聚集标准误差选项，在这种情况下，聚集由个人来加以设定。

一种更有效的估计可利用并入序列相关的模型来获得。经济计量学家强调随机效应。几种软件利用高斯积分法去掉该效应以及在解析形式上易于处理的更特殊随机效应，计数数据模型对含有服从正态分布随机效应的模型进行拟合。不过，统计学家则强调广义线性模型的广义估计方程方法，许多统计软件包与一些经济计量软件均有这些内容可以利用。

若随机效应与回归元相关，则上述这些方法就得出非一致估计。因此，经济计量学家着重固定效应方法。由于有非主要参数问题，所以只对非线性模型的一个子集在短面板情况下的固定效应方法才能得到一致估计。经济计量软件包均有这

些模型的条件极大似然估计，固定效应 logit 与固定效应计数模型可以利用，倘若不能实行固定效应模型，则要使用比最简单的 iid 随机效应模型更为丰富的随机效应模型。

一些动态面板模型也可以被估计出来。这些动态面板模型使得区分由不可观测异质性引起的持久性与由真实状态相依性引起的持久性成为可能。具体执行时，要求编辑各自的计算程序。

23.10 文献注释

本章给出一个略过许多细节的大量且有观点分歧的文献综述。面板数据方面的专著包括阿雷拉诺(Arellano, 2004)、巴尔塔基(Baltagi, 2001)、萧政(Hsiao, 2003)以及李明宰(M.-J. Lee, 2002)，这些文献都对二值数据、删失模型以及选择模型的面板模型进行大量研究。卡梅伦和特里维迪(Cameron and Trivedi, 1998)与李明宰(M.-J. Lee, 2002)的著作均论述了计数数据的面板模型。伍德里奇(Wooldridge, 2002)著作阐述二值数据、删失数据以及计数数据方面的面板方法。各种广义线性模型的统计文献则由法尔迈尔和图茨(Fahrmeier and Tutz, 1994)、迪格尔等人(Diggl et al., 1994, 2002)加以概述。马加什和塞韦斯特(Mátyás and Sevestre, 1995)书中的各篇论文考察非线性面板模型。李明宰(M.-J. Lee, 2002)则着重讨论广义矩方法估计。阿雷拉诺和奥诺雷(Arellano and Honore, 2001)强调了非线性面板模型的半参数方法。库普(Koop, 2003)的著作论述了面板数据的贝叶斯估计。

23.2 对非主要参数问题的一般性讨论，参见兰开斯特(Lancaster, 2002)。关于条件极大似然法的重要考察文献是安德森(Andersen, 1970)，而关于差分法的重要参考文献是张伯伦(Chamberlain, 1992)与伍德里奇(Wooldridge, 1997a)。对于随机效应模型，巴特勒和莫菲特(Butler and Moffitt, 1982)详细阐述了运用高斯求积法剔除服从正态分布的随机效应，不过，统计学参考文献却关注梁和塞格尔(Liang and Zeger, 1986)的广义估计方程法。

23.4 对于固定效应 logit 模型，有关静态模型的重要参考文献是张伯伦(Chamberlain, 1980)，关于纯时间序列的动态模型的重要参考文献则是张伯伦(Chamberlain, 1985)，而含有额外回归元的动态模型方面的参考文献是奥诺雷和基里亚齐杜(Honore and Kyriazidou, 2000)。也可参见萧政(Hsiao, 1995)。

23.5 面板数据选择方面的内容，参见由维拉(Vella, 1998)给出的一个综述，以及巴尔塔基(Baltagi, 2001)与伍德里奇(Wooldridge, 2002)的著作。

23.6 张伯伦(Chamberlain, 1985)论述了各种持续期限模型剔除固定效应的几种方法。范登堡(Van den Berg, 2001, 第 6 节)给出一个优秀的讨论，以及许多参考文献。利用个体多重时期数据的事件历史分析，比大部分面板分析都更为复杂，因为事件历史分析所用模型均是具有内生动态特性。

23.7 面板计数数据模型的经典参考文献是豪斯曼等人(Hausman et al., 1984)的论文。对于动态模型的参考文献，参见布伦德尔等人(Blundell et al.,

2002)的论文。

23.8 面板数据半参数方法的综述,参见阿雷拉诺和奥诺雷(Arellano and Honore, 2001),也可参见李明宰(L.-F. Lee, 2002)。

习 题

23-1 考察非线性面板数据模型 $y_{it}=\alpha_i+exp(\mathbf{x}'_{it}\boldsymbol{\beta})+u_{it}$,其中,$\boldsymbol{\beta}$ 表示待估参数,α_i 表示特定个体效应,$i=1,\cdots,N$,u_{it}是服从 iid $[0,\sigma_\varepsilon^2]$的误差,并且面板数据是短的。

(a) 假定所有 $\alpha_i=0$,能一致估计出 $\boldsymbol{\beta}$ 吗? 若能,请给出该公式或一致估计量的目标函数。若不能,请给出一个简略解释,说明 $\boldsymbol{\beta}$ 为什么不能被一致估计出来。

(b) 假定特定个体效应 α_i 都是随机的,且是 iid $[0,\sigma_\varepsilon^2]$独立服从回归元的。能一致估计出 $\boldsymbol{\beta}$ 吗? 若能,请写出该公式一致估计量的目标函数。若不能,请给出一个简略解释,说明 $\boldsymbol{\beta}$ 为什么不能被一致估计出来。

(c) 假定特定个体效应 α_i 都是随机的,但与回归元是相关的。能一致估计出 $\boldsymbol{\beta}$ 吗? 若能,请写出该公式或一致估计量的目标函数,若不能,请给出一个简略解释,说明 $\boldsymbol{\beta}$ 为什么不能被一致估计出来。

23-2 [改编自张伯伦(Chamberlain, 1980)。]证明,就简单 $T=2$ 模型中 2β 的 plim 而言,二值 logit 面板模型的极大似然估计是非一致的。

23-3 利用与 23.3 节相同的关于专利研发数据的模型,只是因变量与模型要随着下述内容而变动。对于每一种情况,估计随机效应模型。假如从理论上看可行,再估计固定效应模型。

(a) 使用厂商是否有专利的 logit 模型。

(b) 去掉厂商零专利数的观测值,使用对数(专利)数目的截尾 Tobit 模型。

(c) 对专利数使用泊松模型。

第六部分
深入专题

在经验研究中，数据往往不止出现一种情况，而是出现多种需要同时研究的复杂情况。这类复杂问题例子，包括违背简单随机抽样、观测值聚集(clustering，又称聚集、集群)、测量误差以及缺失数据。当它们单独出现或同时出现时，在第四部分与第五部分发展起来的任何模型背景下，关注参数的识别受到损害。第六部分包含三章内容——第 24 章、第 26 章以及第 27 章，分析了此类复杂问题的后果，然后阐述控制这些复杂情况的方法。运用源自本书前面部分的例子阐明方法。这种特性给出第六部分与本书其余部分之间的衔接要点。

第 24 章讨论源自复杂调查数据的几种特征、最著名的分层抽样以及聚集，这也是对第 3 章、第 5 章以及第 16 章所涵盖的各种专题进行一个补充。第 26 章讨论第 4 章、第 14 章以及第 20 章曾经研究过的模型测量误差。第 27 章是关于缺失数据与多重估算的独立一章，但本章利用 EM 算法及吉布斯抽样器，并给出分别与第 10 章及第 13 章的联系特征。

第 25 章阐述处理评估。评估一个宽泛术语，这里评估意指一个变量诸如受教育对某些结果变量诸如工资的影响。处理变量可能是外生指派的，也可能是内生选取的。处理评估专题包括处理对结果影响的识别性，因为处理对结果影响或者通过边际效应进行测算，或者通过边际效应的某些函数进行测算。这里运用一系列方法，包括工具变量回归与倾向得分匹配。处理评估问题，可在第四部分与第五部分考察的任何模型背景下产生。本章强调线性回归模型，故可尽早阅读学习。不过，本章假定读者已熟悉本书涵盖的许多其他专题，包括工具变量与选择模型。因此，我们将日益重要的该专题安排在本书最后部分。

24

分层样本与整群样本

24.1 引 论

通常，微观经济计量学研究通常是借助于人们对关注总体的某个调查样本所搜集的数据来实施的。对调查数据做出的一个最简单统计假设是简单随机抽样(**simple random sampling**，记为 SRS)，在该假设下，总体中的每个元素具有均等概率进入样本中。于是，有理由认为，统计推断建立在数据(y_i，$\mathbf{x}_i$)对不同 i 来说是独立的且服从同分布的假设基础上。这个假设支持了本书阐述的估计量小样本性质与渐近性质，一个显著例外则是第 16 章的样本选择模型。

不过，在实际应用中，简单随机抽样对调查数据来说几乎永远不是一个正确的假设。然而，有一些可供选择的抽样方案，针对人们特别关注总体的子组，以此减少调查成本并增加估计精度。

例如，住户调查可以首先将总体在地理上划分成若干子组，诸如乡村或郊区，然后对各个不同子组执行不同抽样率的调查。访谈可以对那些聚集在小地理区域比如城市街区住户进行。很明显，数据(y_i，$\mathbf{x}_i$)不再服从 iid 的。第一，(y_i，$\mathbf{x}_i$)的分布会随子组不同而变化，故同分布假设可能不适宜。第二，就位于同一群体[1](cluster)的住户而言，数据可能是相关的，所以该群体中(y_i，$\mathbf{x}_i$)是独立的假设失效。

因此，为了获得估计量的分布，就必须对所用的通常方法加以修改，而且估计量的性质可能偏离在简单随机抽样下得到的结果。对此类内容的讨论构成本章主题。

关于回归建模的一些结果如下：第一，若分析目标是对总体特性进行预测，则针对不同抽样率必须加以调整，从而得到加权估计量(**weighted estimators**)。第二，若关注内容在于 y 对 $\mathbf{x}$ 回归，假如给定 $\mathbf{x}$ 时 y 的条件模型得以正确设定，同时分层不是针对因变量的，那么就没有必要实施这类加权。第三，倘若样本借助于因变量的值部分被确定，比如当收入是因变量时，低收入人员的过度样本必须进行加权估

---

[1] 这里翻译成群体或群，意指地理区域位于同一个划分组的众多单元。这样，很容易与“聚集或聚集”(clustering，该词含有动词之意)区分开来。——译者注。

计。此时，有许多估计方法都是可行的，包括第 16 章在样本选择偏倚背景下阐述的某些方法。第四，集群至少会导致标准误差估计相当程度地低估真实标准误差，而且甚至导致非一致参数估计，除非利用类似于第 21 章面板数据分析所阐述的那些方法对集群加以调整。

利用调查数据的大多数微观经济计量学应用都会涉及一项最重要的内容，即需要对聚集加以控制。观测值出现聚集，既时常出现在横截面数据中，又常常出现在面板数据中，其原因有下述三种情况：(1) 抽样设计；(2) 社会实验的设计；(3) 观测方法的性质。情况(1)中的一个例子是复杂大规模住户调查(**complex large-scale household survey**)，为了减少调查成本，对住户某些空间聚集进行抽样。情况(2)中的一个例子是随机化社会实验，将某种共同处理指定给位于特殊位置比如工厂或学校的个体。情况(3)中的一个例子是含有个体横截数据，此时回归元还包括组均值诸如在某个州的失业率或税率、面板数据的运用、双胞胎数据的运用，虽然没有出现住户聚集。

24.2 节介绍抽样调查的一些概念与术语。24.3～24.5 节分别考察调查数据的三个重要特性：样本权重、分层以及聚集。24.6 节考察既出现分层又有聚集情况的分层线性模型。24.7 节讨论数据应用。对复杂调查的进一步研究则由 24.8 节给出。

24.2 抽样调查

在统计学文献中，抽样调查已经得到很好的探索，这是因为数据收集必须在任何分析之前完成，执行调查时其成本费用可能极为昂贵。调查文献的目的通常是以最小成本获得一个样本，该样本能提供总体参数，尤其是总体均值的无偏且合理又准确的估计值。

多阶段调查结构已由 3.2 节描述。美国当前人口调查(CPS)是这类样本设计的一个重要例子。

24.2.1 *当前人口调查*

当前人口调查(Current Population Survey，记为 CPS)是一种每个月大致对 56 000 个住户进行的调查，其目的是作为 16 岁及更大年龄百姓的非公共机构总体的一个代表。在较小州里，住户被过度抽样，以便提供更可靠的州层面数据。为了减少访问成本，某个州内的调查住户被聚集起来。具体地讲，一些住户被连续访问 4 个月，访问停止 8 个月，然后对另一些住户访问。重新访问可减少调查成本，4-8-4 方案容许进行某种纵向分析，包括一年的差异。存在类似大小的 8 个轮换组(**rotation groups**)，每个月都引进一个新轮换组。我们考察有一个轮换组的抽样设计。

具体地讲，存在 792 个层，每一个层是某个州的一个子地区，或者在某些情况下是一个州。792 个层被划分成 2 007 个 PSU，其中，PSU 可能是城市统计区(MSA)，当 MSA 覆盖一个以上州、单个乡村或者两个或更多相邻乡村时，就出现

州与 MSA 交叉，而当 PSU 具有低人口数或大区域时，便违背这个方案。平均地讲，每一个层有 2.5 个 PSU。就 792 个层而言，432 个层才仅有 1 个 PSU，在此情况下，PSU 称为自代表(**self-representing**)，而且总是被包括在 CPS 调查之中。其他 360 个层拥有不到一个 PSU，准确地讲，每一个 PSU 都是随机地从层中选择出来，其概率与 1990 年人口成比例。

对于 PSU 内部来说，不存在中间的 SSU。调查直接对样本 USU 进行抽样，从地区上看，大致有四个地址的紧密组。倘若从层中抽取 PSU 的概率小，则抽样概率增大，而当 PSU 位于小州时，通常会增大抽样概率，允许对人口数少的州过度抽样。(在这种计算中，将纽约和洛杉矶处理成州。)USU 中的全部住户都会被调查，除非 USU 拥有异常多的住户，在此情况下，就要对住户子集进行随机抽取。

CPS 被设计成利用州自加权(**self-weighting**)，因而尽管使用非随机抽样，但 CPS 应该为每一个州提供一个代表性样本。不过，未加权的样本当然不是代表性的，因为对人口数少的州进行过度抽样，并且不是所有的 PSU 都能被抽取到。

24.2.2 *抽样*

在离开抽样调查的更详细分析之前，我们在没有复杂情况比如分层情况时对抽样基础给出一个简略描述。

设 $\mathbf{z}$ 表示变量向量，这里不必对因变量与回归元变量加以区分。我们假定，总体中变量 $\mathbf{z}$ 是 iid 的，具有密度 $f(\mathbf{z})$。总体是一个容量为 N^* 的总体，而样本具有容量 N。样本是$\{\mathbf{z}_i, i=1,\cdots,N\}$，其中，$i$ 表示第 i 个抽样单元。在抽样文献中，通常符号 n 用于表示样本量，而 N 用于表示总体容量。不过，我们继续用 N 表示样本量，因为仅在偶尔机会才引入总体容量 N^*。

穷尽抽样

在穷尽抽样(**exhaustive sampling**)下，总体的每一个元素都会被抽取，因此，样本就是总体。这种抽样在个体层面数据上极少用到。在人口普查诸如美国 10 年一次的人口普查中，就会遇到。然而，甚至对普查来说，子抽样用于较长问卷，研究人员更喜欢用易于管理的普查子样本来展开工作，而且在实际应用中，普查覆盖面是不完整的。对厂商层面数据而言，穷尽抽样则更为普遍采用，例如，某个行业的全部厂商可能都是研究内容。

穷尽抽样引发了对通常推断方法是否合理的争论，因为样本矩等于总体矩。通常程序还是使用通行的推断方法。这样做时，是将有限样本看成来自一个无限超总体(**superpopulation**)的一个样本。

例如，假定关注内容是工作场地中性别之间工资的差异，该工作场地有包含 20 名男性与 12 名女性的一个总体，他们的工作任务类似。对于工作场地上所有男性与女性来说，他们都赚取工资，故这个样本是总体，可以发现，就平均工资而言，男性高于女性。一种习惯做法是，对平均工资差异进行传统假设检验，而不是得出如下结论：由于样本均值等于总体均值，所以有 100%的把握确定，男性工资较高。其根本原因是，在此特定工作场地的总体被看成是来自工作场地超总体的一个样本，或来自在众多时点上特定工作场地超总体的一个样本。

穷尽抽样费用昂贵，而且一般地讲，对大样本来说，不必用穷尽抽样，除非实际总体大小必须是确定的。相反，通常对总体的子集加以抽样。

简单随机抽样

简单随机抽样(**simple random sample**)是指那种观测值随机地从总体抽取且具有同等概率的抽样。样本出现的每一个观测值都具有等于样本量被总体容量除的概率，同时拥有相同的边缘密度 $f(\mathbf{z})$。这里添加定语“简单”，因为更系统的抽样方法通常还具有随机元素。

有限样本校正

大多数经济计量分析假定，SRS 会产生 $\mathbf{z}$ 采样，这些采样是独立的，因此，在 SRS 条件下，样本的联合密度是单个密度 $f(\mathbf{z}_i)$的乘积。假如 SRS 是来自一个无限总体，就如同将抽样看成是来自一个超总体，或来自一个有限总体且抽样是放回的，这样做是有道理的。

在实际应用中，对于有限总体来说，SRS 是不放回的，以此保证同一观测值不会两次出现在样本之中。于是，甚至在 SRS 条件下，观测值不再是独立的，为了理解这一点，注意到，在 SRS 条件下，出现在样本中的总体任意特定元素的概率是 N/N^*。不过，已知此元素出现在样本中，则样本出现的任何其他元素的概率为 $(N-1)/(N^*-1)$。很明显，条件概率不同于无条件概率。更正式地讲，人们引进一个指示变量以表明总体的每一种情况是否出现在样本中。这些指示变量服从联合多项式分布，其均值为 π，方差为 $\pi(1-\pi)$，而协方差为 $-\pi(1-\pi)/(N^*-1)$，其中，$\pi=N/N^*$。

样本观测值之间的相关是 $\rho=-1/(N^*-1)$，其中，ρ 被称为群内相关(**intra-class correlation**)。设 z 是一个纯量，得出该样本均值为 $\bar{z}=N^{-1}\sum_i z_i$，其方差为 $\mathrm{V}[\bar{z}]=N^{-2}\mathrm{V}[\sum_i z_i]$，该式并不能简化成 $N^{-2}\sum_i \mathrm{V}[z_i]$，原因是 z_i 的相关性。例如，经过一些代数运算，克柯伦(Cochran，1977，第 23～24 页)得出：

$$\mathrm{V}[\bar{z}]=(1-f)\frac{S^2}{N}$$

其中，$f=N/N^*$ 表示抽样比例，克柯伦著作中的一些结果经常利用 $S^2=(N^*-1)^{-1}\times\sum(z_i-\bar{z})^2$，而不是利用通常有限总体方差 $\sigma^2=N^{*-1}\sum(z_i-\bar{z})^2$ 加以进一步简化。

因而，对于源自有限总体的不放回抽样来说，样本均值的方差等于通常 S^2/N 乘以一个有限样本校正项(**finite-sample correction term**)$1-f$。该校正项出现在调查数据的统计软件包中，当不考虑有限样本校正项时，就产生传统的统计推断，因为 $\mathrm{V}[\bar{z}]$被过高估计。对于使用来自放回 SRS 数据的回归来说，类似地，有限样本校正是一个有关的内容，尽管 OLS 估计量方差的偏倚范围及程度现在都另外依赖于设计矩阵。

在微观经济计量学中，有限样本校正项经常被忽略掉。这样做，经常是合理的。例如，对于住户调查数据来说，样本量相对于总体容量而言很小，所以 $f=N/N^*\rightarrow 0$。

24.3 加 权

住户调查比如 CPS，经常以下述方式建立起来：不同住户拥有进入样本的不同

概率。为了校正这种情况,对每一个观测值都要指派一个权数。

正如下面将要解释的,倘若分层是外生的,若将回归看成是一个工具,以此刻画总体响应,就应使用权数,可是如果回归模型被假定成一种正确的结构模型,就不需要使用权数。

24.3.1 *样本权数*

假定总体的每一个住户出现在样本 i 中的概率为 π_i,而且假定该概率与 SRS 不同,随着不同住户而变化。

诸如一般样本均值统计量,对所有观测值都给予同等权数,这将出现对以高概率出现在样本中的住户给予太大的权数。可运用加权方式来修正上述情况,即利用与包含于样本的概率成反比例的样本权数:

$$w_i \propto 1/\pi_i \tag{24.1}$$

例如,我们可以用加权均值:

$$\bar{z}_{\mathrm{W}} = N^{-1}\sum_i w_i z_i / \sum_i w_i$$

代替 $\bar{z}=N^{-1}\sum_i z_i$。注意到,式(24.1)的所有问题具有比例性。倘若我们用权数之和去除,则不必要求权数之和为 1。一种共同标度是$\sum_i w_i = N^*$,在此情况下,权数 w_i 意味着该观测值代表总体中 w_i 住户。注意,使用权数时小心慎重。相反,某些推断定义 $w_i \propto \pi_i$,而某些计算机软件中,加权均值作为$\sum_i(z_i/w_i)/\sum_i(1/w_i)$。利用样本权数的倒数,很容易不正确地加权。

对于来自容量为 N^*的有限总体的容量为 N 的 SRS,$\pi_i=1/N^*$,故 w_i 是一个常值,从而 $\bar{z}_{\mathrm{W}}=\bar{z}$。

对于层内具有 SRS 的简单分层抽样,假定知道,总体容量 N^* 的一部分 H_s 处于第 s 层,并且 N_s 个观测值来自第 s 层,那么 $\pi_i=N_s/H_sN^*$。由此可得,样本权数 $w_i \propto H_s/N_s$。

对于两阶段无分层抽样,设 π_c 是第 c 个 PSU 被抽取的概率,而 π_{jc}是位于第 c 个 PSU 之中的住户 j 被抽取的概率。那么样本权数 $w_{jc} \propto 1/(\pi_c N_c \pi_{jc} N)$,其中,$N_c$ 表示位于第 c 个 PSU 中的调查住户数,而 $N=\sum_c N_c$。倘若每一阶段抽样概率均与总体数成比例,则两阶段样本是自加权的,因此 $\pi_c=N_c^*/N^*$,$\pi_{jc}=1/N_c^*$,其中,N_c^* 表示第 c 个 PSU 的总体数。于是,如同 SRS 一样,权数 w_{jc}都是同等的,尽管就两阶段抽样而言,估计量标准误差还必须加以调整,如同 24.8 节将要证明的。

对于 CPS,即对小州住户过度抽样,看起来使用 $w_i \propto H_s/N_s$ 就足够了,其中,s 表示州。CPS 将上式作为一个基准权数,可是当 USU 拥有太多住户时,对 USU 内的二次抽样加以调整;因此,如果抽样 PSU 与其层面标准相差甚远时,此层的调查住户可能不是层的代表。这就导致了两个另外调整。首先,针对层水平的非代表性种族(黑人/非黑人)组成进行调整。其次,为了确保(由州、民族、女性或年龄形成的)重要子组的样本估计值匹配独立总体数据,要对权数加以调整。有关详细内容,参见美国人口普查局(2002)。一旦控制住源自州、民族、性别以及年龄维度

的美国居民总体所引起的 CPS 差异，为了允许 CPS 提供自然代表性统计量，要建立 CPS 样本权数。

对于多阶段调查来说，实际样本权数的计算涉及相当复杂的估计程序。权数可能被错误地估计；即使权数被正确地估计出，但权数可能考虑到唯一的样本非代表性的某些维度。

24.3.2 *加权回归*

当得到样本权数时，人们应该怎样实施加权回归呢？当分层不是针对因变量时，我们详细考察这个问题。

关于因变量的分层，24.4 节将给予讨论。

考察线性回归

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i \tag{24.2}$$

的估计，已知含有抽样权数 w_i 的调查数据。两种可行估计量分别是 OLS：

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{24.3}$$

以及使用抽样权数的 WLS：

$$\hat{\boldsymbol{\beta}}_{WLS} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y} \tag{24.4}$$

其中，$\mathbf{W} = \mathrm{Diag}[w_i]$。

正确设定条件均值

如果假定 $\mathrm{E}[\mathbf{u}|\mathbf{x}] = \mathbf{0}$，那么 OLS 估计量就是合适的，因而其条件均值关于是 $\mathbf{x}$ 线性的：

$$\mathrm{E}[y_i|\mathbf{x}_i] = \mathbf{x}_i'\boldsymbol{\beta} \tag{24.5}$$

于是，OLS 关于 $\boldsymbol{\beta}$ 是一致的。进一步地，如果误差 u_i 都是同方差的，由高斯—马尔可夫定理知，OLS 是二阶矩有效的。在这些假设下，WLS 估计量关于 $\boldsymbol{\beta}$ 也是一致的，但当误差是同方差时[由于式(24.5)中权数控制样本为非代表性而不是异方差性]，WLS 估计量将是非有效的。

不正确设定条件均值

在许多应用中，式(24.5)并不成立。一些例子包括下述情况：省略回归元或当 $\mathrm{E}[y|\mathbf{x}]$ 关于 $\mathbf{x}$ 是非线性的情况，或者 $\mathrm{E}[y_i|\mathbf{x}_i] = \mathbf{x}_i'\boldsymbol{\beta}_i$，其中，$\boldsymbol{\beta}_i$ 的某些成分与 $\mathbf{x}_i$ 相关。线性回归还能被解释成在误差平方损失条件下，给定 $\mathbf{x}$ 时 y 的最佳线性预测，尽管这需要适当考虑到非代表性抽样。

在总体中，$(y_i, \mathbf{x}_i)$ 是 iid 的，并且由 4.2 节知，我们总能写成：

$$y_i = \mathbf{x}_i'\boldsymbol{\beta}^* + u_i$$

其中，$\mathrm{E}[u] = 0$，$\mathrm{Cov}[\mathbf{x}, u] = 0$，而：

$$\boldsymbol{\beta}^* = (\mathrm{E}[\mathbf{x}\mathbf{x}'])^{-1}\mathrm{E}[\mathbf{x}y]$$

注意到，不再假定 $\mathrm{E}[\mathbf{u}|\mathbf{x}] = \mathbf{0}$，因而可能有 $\mathrm{E}[y|\mathbf{x}] \neq \mathbf{x}'\boldsymbol{\beta}$。

杜穆谢尔和邓肯(DuMouchel and Duncan，1983)将参数 β^* 称为普查系数(**census coefficient**)。回归系数的概率极限，可通过进行回归而获得，这里使用了整个总体而不是非代表性样本。

如果条件均值关于 $\mathbf{x}$ 是非线性的，并且样本是总体的非代表性样本，那么一般地讲，OLS 估计量并不收敛到 β^*，因为就非代表性样本而言，$N^{-1}\mathbf{X}'\mathbf{X}$ 不收敛到总体矩 $\mathrm{E}[\mathbf{xx}']$，类似地，$N^{-1}\mathbf{X}'\mathbf{y}$ 也如此。从直观上看，若条件均值关于 $\mathbf{x}$ 是非线性的，则没有理由认为，当运用同一总体的不同调查样本时，线性回归将会得出同样的 OLS 估计值。

不过，运用样本权数的 WLS 可以一致地估计出 β^*。具体地讲，如果加权矩阵 $\mathbf{W}$ 使得：

$$N^{-1}\mathbf{X}'\mathbf{W}\mathbf{X} \xrightarrow{p} \mathrm{E}[\mathbf{xx}'] \tag{24.6}$$

$$N^{-1}\mathbf{X}'\mathbf{W}\mathbf{y} \xrightarrow{p} \mathrm{E}[\mathbf{x}y]$$

那么式(24.4)定义的 $\hat{\beta}_{\mathrm{WLS}}$ 收敛到 β^*。

简单分层样本

加权 LS 估计的绝大多数分析，都是对于在层内满足 SRS 的简单分层抽样阐述的。于是，很显然，如果第 i 个受访住户位于第 s 层，那么式(24.6)就满足 $w_i \propto H_s/N_s$。

这方面的文献还考虑到，层内各种不同回归系数的可能性。假定 $\mathrm{E}[y_i|\mathbf{x}_i]=\mathbf{x}_i'\beta$，对于位于 s 层的住户。其目标或许是估计总体加权参数 $\beta_W=N^{-1}\sum_s N_s^* \beta_s$。那么，一般地讲，不论 OLS 还是 WLS 都不收敛到 β_W，除非 β_s 对不同层而言是同等的或者是 iid 的，具有常值均值。该结果的一个著名例外是，对 y 均值的估计(即 $\mathbf{x}=1$ 时的回归)，在此情况下，层样本均值的加权平均关于总体均值是无偏的。对于详细内容，参见 24.4.1 节以及杜穆谢尔和邓肯(DuMouchel and Duncan，1983)、迪顿(Deaton，1997)或者乌拉和布罗伊尼希(Ullah and Breunig，1998)。

人们应该用样本权数吗?

上述分析能用于回答，一旦假定不存在内生分层时，是否将样本权数用于估计中。此处讨论考察 $\mathrm{E}[y|\mathbf{x}]$ 模型的(可能非线性)估计问题，但是也可应用于给定 $\mathbf{x}$ 时 y 的条件分布的任何其他特定上，诸如中位数或密度。

假如人们采用结构方法或解析方法，并假定 $\mathrm{E}[y|\mathbf{x}]$ 模型得以正确设定，则不必使用样本权数。其结果能用于分析 $\mathbf{x}$ 变化时 $\mathrm{E}[y|\mathbf{x}]$ 的效应。

相反，假如人们采用描述方法或数据汇总方法，则应使用权数。于是，回归被解释成估计普查系数。不过，一个重要告诫是，在复杂调查中，很明显不可能估计满足式(24.6)的权数，因为这是层内满足 SRS 的分层抽样情况。在实际应用时，针对基于年龄、性别与民族的某些子组，建立抽样权数匹配总体比例。无法保证使得此类权数满足式(24.6)。

对于某些数据集，比如有几千个住户小的纵向调查，的确要发展一种结构建模方法。不过，此类数据集通常试图提供总体的一个合理代表性样本，利用整群抽样缩减调查成本。其他一些数据集，例如，CPS 被设计成能提供准确描述测量，比如

国家与地区的失业率估计。这里，调查设计者采用了普查方法，并且事实上当每月实施普查费用并不昂贵时，更喜欢每月实施普查。

对于上述两种数据集的任何一类，微观经济计量学家通常力争采用结构建模方法(**structured modeling approach**)。举一个例子，考察收入对受教育水平与社会经济特征——比如年龄、性别以及民族，却没有固有能力——的测量。

大多数经济计量学家愿意对 OLS 回归中受教育系数给出一个描述性解释，原因在于受教育的内生性。于是，其解释如下：如果我们保持某些重要的回归为恒定，那么多受一年教育会引起收入增加 6%的相当变化，但不一定是因果关系。这里，OLS 回归中的样本权数适合于允许将估计解释成对总体中相连部分而不是那些仅仅可能非代表性样本的测量。即使因果解释不可行，这个估计值也是有用的。因为它测算出一旦控制住某些其他重要的社会经济变量后，对于受教育不同组别而言收入是如何变化的。统计量的主要目的毕竟是数据汇总。

受教育系数的一致估计值，可通过利用更高等的估计方法比如工具变量或面板数据方法来获得。于是，该系数就能给出一种因果解释。借助于样本权数进行加权不再是必需的，尽管如果误差是异方差的，通常加权会改进效率。

一个模型能否解释成被正确设定，这是一种主观判断。如果模型被正确设定，那么样本加权与不加权的估计应具有相同的概率极限，两者的这两个估计都是一致的。这就建议，利用样本加权估计量与不加权样本估计量之差的豪斯曼检验对正确模型设定加以检验，这种检验是由杜穆谢尔和邓肯(DuMouchel and Duncan, 1983)在线性回归情况下提出的。

24.3.3 *预测*

考察具有正确设定条件均值 $g(\mathbf{x},\boldsymbol{\beta})$且无内生性的非线性回归。非加权 NLS 估计量一致地估计出，并且给出了因果解释。特别地，我们能使用$\partial g(\mathbf{x},\hat{\boldsymbol{\beta}})/\partial\mathbf{x}$ 计算当 $\mathbf{x}$ 变化一个单位时引起条件均值的因果效应。

由于 $g(\cdot)$是非线性的，所以预测效应会随着计算点 $\mathbf{x}$ 不同而变化。总体平均响应的一个估计是：

$$\hat{\mathrm{E}}\left[\frac{\partial y}{\partial \mathbf{x}}\right]=\sum_{i=1}^{N} w_i \frac{\partial g(\mathbf{x}_i,\hat{\boldsymbol{\beta}})}{\partial \mathbf{x}_i}$$

其中，w_i 表示样本权数。类似地，如果相反人们在回归元均值处计算其响应，那么使用 $\mathbf{x}$ 的总体均值估计值，即 $\mathbf{x}$ 的加权样本均值而不是 $\mathbf{x}$ 的非加权样本均值会更好。

即使通过利用非加权估计能一致地估计出参数，可是假若人们想要预测总体影响而不是样本影响，则在后面计算影响时必须运用加权。

24.4 内生分层

分层被人们广泛地使用，因为它能提高估计准确性，或等价地讲，在给定准确性水平时能减少调查成本。例如，在人口少的州里，平均失业率更为准确的估计可

通过对该州过度抽样来获得。缘于类似的原因，对少数组可进行过度抽样。

一种新的困难是，参数会随不同州而变化，这一点已在 24.3 节考虑过。例如，平均失业率可能随不同州而变化。于是，就要采用描述方法，并使用加权估计量。

微观经济计量学家经常更愿意采用结构方法，同时假定参数对不同州而言均为常值。那么，由 24.3 节知，分层很明显并不会引起新的困难，从而使用非加权回归。一个重要条件是，如果分层建立在因变量值基础上，那么还是出现了问题。例如，若低收入人员被故意地过度抽样，并且收入作为因变量，则通常回归估计量都是非一致的。注意到，假如分层是关于回归元的，比如性别，则不会产生问题，这导致间接地对低收入人员的过度抽样。如果分层是直接针对收入的，才会出现问题。

本节我们将定义内生分层，并分析其后产生的复杂问题。然后，我们阐述几个一致估计量。最简单的是一种加权估计量，若既已知样本分层概率又已知总体分层概率，就可运用该估计量。这种方法由 24.4.5 节给出，以自给自足方式加以阐述。

24.4.1 *分层方案*

对于一般数据 $\mathbf{z}\in\mathcal{Z}$，层是 $\mathcal{Z}$ 的一些子集。经济计量分析通常将数据分割成因变量 $\mathbf{y}\in\mathcal{Y}$ 与回归元或自变量 $\mathbf{x}\in\mathcal{X}$，其中，考虑到一般性，我们允许 $\mathbf{y}$ 是一个向量。于是，层 $\mathcal{C}_s$ 被定义成样本空间 $\mathcal{Y}\times\mathcal{X}$ 的子集，$s=1,\cdots,S$。这种记号由英伯斯和兰开斯特(Imbens and Lancaster, 1996)使用过，他们阐述了某些重要例子，表 24.1 重新给出其内容。

表 24.1　层内为随机抽样分层方案

分层方案	定　　义	对层的描述
简单随机抽样	$S=1$，$\mathcal{C}_1=\mathcal{Y}\times\mathcal{X}$	一个层涵盖整个样本空间
纯外生分层方案	$\mathcal{C}_s=\mathcal{Y}\times\mathcal{X}_s$，满足 $\mathcal{X}_s\subset\mathcal{X}$	仅以回归元而不以因变量进行分层
纯内生分层方案	$\mathcal{C}_s=\mathcal{Y}_s\times\mathcal{X}$，满足 $\mathcal{Y}_s\subset\mathcal{Y}$	仅以因变量而不以回归元进行分层
增广样本	$S=2$，$\mathcal{C}_1=\mathcal{Y}\times\mathcal{X}$，且 $\mathcal{C}_2\subset\mathcal{Y}\times\mathcal{X}$	通过样本空间的额外观测值进行增广随机样本
分割	$\mathcal{C}_s\subset\mathcal{Y}\times\mathcal{X}$，$\mathcal{C}_s\cap\mathcal{C}_t=\varnothing$ 且 $\bigcup_{s=1}^{S}\mathcal{C}_s=\mathcal{Y}_s\times\mathcal{X}$	样本空间被分割成互斥且充满整个样本空间的众多层

层内抽样被假定成随机的，只是某些层可能被过度抽样。由表 24.1 可知，很明显，各层之和可以小于或大于样本空间。对于第四个方案与第五个方案，分层可能仅仅针对内生变量、外生变量，或者针对这两者的混合。

经济计量学文献关于含有内生成分的抽样方案，因为在这种情况下，通常条件 MLE 是非一致的。

内生分层已在第 16 章讨论过。举一个例子，考察截尾回归，其中，只有当$y>0$

时我们才观测到 y,所以分层是完全针对 y 的。那么,对于抽样数据来说,给定 $\mathbf{x}$ 时 y 的条件密度是零截尾密度,它是用 $\Pr[y>0|\mathbf{x}]$ 去除非截尾密度,因而:

$$f^s(y|\mathbf{x},\boldsymbol{\theta})=\frac{f(y|\mathbf{x},\boldsymbol{\theta})}{1-F(0|\mathbf{x},\boldsymbol{\theta})}$$

其中,上标用于区分总体密度 $f(y|\mathbf{x},\boldsymbol{\theta})$ 与样本密度。正如第 16 章讨论的,这个抽样方案倾向于略去给定 $\mathbf{x}$ 时具有 y 的低水平实现值的观测值。假定 $\mathrm{E}[y|x]=\beta_1+\beta_2 x$,并且 $\beta_2>0$。于是,对于 x 的低水平值,存在极多 y 的相对高水平值。因此,就 x 的低水平值而言,该回归将过度预测 $\mathrm{E}[y|x]$,导致截距 β_1 出现向上偏倚,斜率 β_2 出现向下偏倚。

第二个例子是二值数据或多项式数据的基于选择抽样,其中,样本是基于离散结果 y 来加以选取。例如,若在乘大巴与乘小车往返工作之间进行选择,我们可能过度抽样那些相对少数的大巴乘客。该例子以下述方式继续探讨。它类似于医学文献中的病例对照研究(**case-control studies**),例如,因某种疾病而死亡($y=1$)的人员的一个完整样本与因该种疾病而未死亡的人员($y=0$)全体的一个大小类似的子样本加以对照。其目标是找到不止一个回归元能否预测 $y=1$。

一个有关的例子是,通过用户现场抽样(**on-site sampling**)搜集到的访问数目的计数数据,诸如娱乐场合或购物中心或者医生办公室。于是,数据被截尾,因为满足 $y=0$ 的那些没有被抽取,另外高额访问者被过度抽样。邵(Shaw,1988)已经证明,数据抽样分布 $f^s(y|\mathbf{x},\boldsymbol{\theta})$ 通过方程

$$f^s(y|\mathbf{x},\boldsymbol{\theta})=f(y|\mathbf{x},\boldsymbol{\theta})\frac{y}{\mathrm{E}[y|\mathbf{x},\boldsymbol{\theta}]}$$

与总体分布相联系,在此情况下,很明显,该抽样方案是内生的,

24.4.2 *分层诱导内生性*

抽样方案诸如分层方案导致样本密度不同于总体密度。假如分层仅仅是外生的,则尽管有这样的差异,但就样本而言,由于给定的条件密度与其总体的一样,所以通常仍是一致的。不过,如果分层的任何方面都是内生的,那么正如上述例子所阐明的,这些条件密度会不一样。现在,我们对该问题给予详细讨论。

ML 估计的目的在于一致地估计出 $f(y|\mathbf{x},\boldsymbol{\theta})$ 中的参数 $\boldsymbol{\theta}$。一般地讲,MLE 应建立在来自数据($y,\mathbf{x}$)的联合分布的似然函数基础上。实际上,直接从源自给定 $\mathbf{x}$ 时 y 的条件分布建立起条件似然函数,这样做时常就足够了。这种较简单方法在下面假设下能产生一致估计,该假设是:对于 y 来说,$\mathbf{x}$ 是外生的,在此情况下,其联合密度分解成:

$$g(y,\mathbf{x}|\boldsymbol{\theta})=f(y|\mathbf{x},\boldsymbol{\theta})\times h(\mathbf{x}|\boldsymbol{\theta}) \tag{24.7}$$

其中,$\mathbf{x}$ 的密度参数没有标出来,因为人们不希望估计这些参数。

我们总是能写成 $g(y,\mathbf{x})=f(y|\mathbf{x})\times h(\mathbf{x})$。式(24.7)做出的假设是,引入参数 $\boldsymbol{\theta}$,$\boldsymbol{\theta}$ 出现在 $f(y|\mathbf{x},\boldsymbol{\theta})$ 之中而不出现 $h(\mathbf{x})$ 中。一般地讲,我们更愿意写成:

$$g(y,\mathbf{x}|\boldsymbol{\theta})=f(y|\mathbf{x},\boldsymbol{\theta})\times h(\mathbf{x}|\boldsymbol{\theta}) \tag{24.8}$$

而不是写成式(24.7)。就 y 而言，$\mathbf{x}$ 的1个或更多成分是内生的，因为现在存在着一种反馈，y 依赖于 $\mathbf{x}$，但 $\mathbf{x}$ 通过在 $h(\mathbf{x}|\boldsymbol{\theta})$ 中出现 $\boldsymbol{\theta}$ 而反过来依赖于 y。这方面的一个经典例子是线性联立方程。在此类情况下，ML 估计应建立在联合似然函数

$$\ln L_{\text{JOINT}}(\boldsymbol{\theta}) = \sum_{i=1}^{n}\ln f(y_i|\mathbf{x}_i,\boldsymbol{\theta}) + \sum_{i=1}^{n}\ln h(\mathbf{x}_i|\boldsymbol{\theta}) \tag{24.9}$$

的基础上。由第1章知道，如果：

$$\mathbf{0}=\mathrm{E}\left[\frac{\partial\ln g(y,\mathbf{x}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right]=\mathrm{E}\left[\frac{\partial\ln f(y|\mathbf{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right]+\mathrm{E}\left[\frac{\partial\ln h(\mathbf{x}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right] \tag{24.10}$$

就可得到 $\boldsymbol{\theta}$ 的一致估计。当密度 $g(y,\mathbf{x}|\boldsymbol{\theta})$ 得以正确设定且数据范围不依赖于 $\boldsymbol{\theta}$，条件(24.10)得以满足。不过，条件 MLE 极大化条件似然函数：

$$\ln L_{\text{条件}}(\boldsymbol{\theta}) = \sum_{i}\ln f(y_i|\mathbf{x}_i,\boldsymbol{\theta})$$

当 $\mathrm{E}[\partial\ln f(y|\mathbf{x},\boldsymbol{\theta})/\partial\boldsymbol{\theta}]=\mathbf{0}$ 时，条件 MLE 是一致的。倘若 $\mathbf{x}$ 是外生的，由于 $\partial\ln h(\mathbf{x})/\partial\boldsymbol{\theta}=\mathbf{0}$，故式(24.10)可以简化，那么式(24.10)意味着这个必要条件成立。相反，若 $\mathbf{x}$ 是内生的，则这种简化就不会出现，因为式(24.10)右边的第二项不会消失。因而，当 $\mathbf{x}$ 是内生的时候，条件 MLE 是非一致的。

分层以及类似的抽样方案所产生的问题是，即使总体联合密度满足式(24.7)且对不同层来说都是相同的，抽样方案能导致样本中的 $(y,\mathbf{x})$ 联合密度采取更为一般的形式：

$$g^s(y,\mathbf{x}|\boldsymbol{\theta})=f^s(y|\mathbf{x},\boldsymbol{\theta})\times h^s(\mathbf{x}|\boldsymbol{\theta}) \tag{24.11}$$

其中，上标"s"用于表示对所用特殊抽样方案的依赖性。那么，条件 MLE 可能是非一致的，尽管若样本是 SRS 时，条件 MLE 会是一致的。

在纯外生抽样(**pure exogenous sampling**)条件下，对于 $\mathbf{x}$ 的边缘密度来说，样本分布与总体分布之间出现唯一的差异。假定就总体而言，式(24.7)成立，则样本有：

$$g^s(y,\mathbf{x}|\boldsymbol{\theta})=f(y|\mathbf{x},\boldsymbol{\theta})\times h^s(\mathbf{x})$$

很明显，条件 MLE 将是一致的，因为条件密度仍然是 $f(y|\mathbf{x},\boldsymbol{\theta})$，并且 $\boldsymbol{\theta}$ 没有出现在 $h^s(\mathbf{x})$ 中。

在内生抽样方案下，显然就总体而言，式(24.7)成立，但作为更一般结果，式(24.11)对样本来说成立。给定 $\mathbf{x}$ 时，y 的样本条件分布与总体条件分布可以不同，有 $f^s(y|\mathbf{x},\boldsymbol{\theta})\neq f(y|\mathbf{x},\boldsymbol{\theta})$，并且 $h^s(\mathbf{x}|\boldsymbol{\theta})$ 可能依赖于 $\boldsymbol{\theta}$。

24.4.3 *内生抽样*

在纯内生抽样下，样本 y 的边缘分布不同于其总体边缘分布。设 $h(y)$ 表示 y 的总体密度，而 $h^s(y)$ 表示 y 的抽样密度。[我们依照惯例，用 g、f 以及 h 分别表示联合、条件以及边缘分布。很明显，读者会分辨 $h(y)$ 与 $h(\mathbf{x})$。]

在纯内生抽样下，y 与 $\mathbf{x}$ 的联合分布最好是通过先以 $\mathbf{x}$ 而不是 y 为条件来获得。于是：

$$g^s(y,\mathbf{x})=f(\mathbf{x}|y)h^s(y) \tag{24.12}$$

其中，给定 y 时 $\mathbf{x}$ 的条件分布在纯内生抽样下并不会受到影响，故出现简化，从而 $f^s(\mathbf{x}|y)=f(\mathbf{x}|y)$。现在，我们需要用 $f(\mathbf{x}|y)$ 重新表述出 $f(y|\mathbf{x})$。从而：

$$\begin{aligned} f(\mathbf{x}|y)&=\frac{g(y,\mathbf{x})}{h(y)} \\ &=\frac{f(y|\mathbf{x})h(\mathbf{x})}{h(y)} \end{aligned} \tag{24.13}$$

将式(24.13)代入式(24.12)，并重新整理得到：

$$g^s(y,\mathbf{x}|\boldsymbol{\theta})=f(y|\mathbf{x},\boldsymbol{\theta})\times\frac{h^s(y)}{h(y|\boldsymbol{\theta})}\times h(\mathbf{x})$$

其中：

$$\begin{aligned} h(y|\boldsymbol{\theta}) &= \int g(y,\mathbf{x}|\boldsymbol{\theta})d\mathbf{x} \\ &= \int f(y|\mathbf{x},\boldsymbol{\theta})h(\mathbf{x})d\mathbf{x} \end{aligned}$$

仅利用 $f(y|\mathbf{x},\boldsymbol{\theta})$ 的条件 MLE 将是非一致的，因为 $h(y|\boldsymbol{\theta})$ 项可被忽略掉。然而，人们需要对另外包括 $h(y|\boldsymbol{\theta})$ 的联合似然求极大值。

24.4.4 *内生分层样本*

我们现在考察 24.4.1 节已经进入的分层方案。其总体密度是：

$$g(y|\mathbf{x},\boldsymbol{\theta})=f(y|\mathbf{x},\boldsymbol{\theta})h(\mathbf{x})$$

这里，存在 S 个层，第 s 层是 $\mathcal{Y}\times\mathcal{X}$ 的子集 C_s。

位于 C_s 内观测值的总体概率与源自 C_s 的抽样概率之间有重要差异，因为两者在分层抽样方案上不同。我们定义：

$$\begin{aligned} H_s&=\Pr[\text{从 } C_s \text{ 中抽取到的观测值}] \\ Q_s(\boldsymbol{\theta})&=\Pr[\text{从由 } C_s \text{ 构成的总体中随机抽取观测值}] \end{aligned} \tag{24.14}$$

这里，H_s 是借助于样本设计的集合，而：

$$Q_s(\boldsymbol{\theta}) = \int_{C_s} f(y|\mathbf{x},\boldsymbol{\theta})h(\mathbf{x})dyd\mathbf{x} \tag{24.15}$$

层概率可能是未知的，也可能不是未知的。当 $H_s>Q_s$ 时，就出现对层过度抽样。

我们通过获得 s、y 以及 $\mathbf{x}$ 的联合密度来开始讨论，其中，s 是一个指示变量，表示观测值是来自哪一层。在总体中：

$$g(s,y,\mathbf{x}|\boldsymbol{\theta})=Q_s(\boldsymbol{\theta})g(y,\mathbf{x}|s,\boldsymbol{\theta})$$

就样本而言，层指示变量的边缘分布不同于 Q_s，并且：

$$
\begin{aligned}
g^{s}(s,y,\mathbf{x}|\boldsymbol{\theta}) &= H_{s}g(y,\mathbf{x}|s,\boldsymbol{\theta}) \\
&= H_{s}\frac{f(y|\mathbf{x},\boldsymbol{\theta})h(\mathbf{x})}{Q_{s}(\boldsymbol{\theta})}
\end{aligned}
$$

其中,第二个等式成立,这是因为 $g(y,\mathbf{x}|s)$ 等于密度 $g(y,\mathbf{x})=f(y|\mathbf{x})h(\mathbf{x})$ 被位于 s 层的总体概率除,从而在 C_s 上的积分为 1。

由此可得,其联合密度是:

$$
g^{s}(s,y,\mathbf{x}|\boldsymbol{\theta})=\frac{H_{s}}{Q_{s}(\boldsymbol{\theta})}f(y|\mathbf{x},\boldsymbol{\theta})h(\mathbf{x}) \tag{24.16}
$$

其中,$Q_s(\boldsymbol{\theta})$ 已由式(24.15)定义。基于总体条件密度 $f(y|\mathbf{x},\boldsymbol{\theta})$ 的条件 MLE 关于 $\boldsymbol{\theta}$ 是非一致的,原因在于它忽略了依赖于 $\boldsymbol{\theta}$ 的项 $Q_s(\boldsymbol{\theta})$。

人们可提出一系列的一致估计量。此处,我们考察极大似然估计、GMM 估计以及更为简单的加权估计量,这类估计量实施起来既能提供层抽样概率 H_s,还可以知道总体概率 $Q_s(\boldsymbol{\theta})$。

极大似然估计

实施基于式(24.16)中联合密度 $g^{s}(s,y,\mathbf{x}|\boldsymbol{\theta})$ 的 ML 估计是一项复杂任务,因为由式(24.15)知,$Q_s(\boldsymbol{\theta})$ 的分布依赖于 $h(\mathbf{x})$。一种可行求解是设定密度 $h(\mathbf{x})$。该方法并没有被采用,因为经济计量学家要避开设定回归元分布,尽管有设定因变量条件分布的意愿。

相反,对于未设定 $h(\mathbf{x})$,人们采用半参数方法,其目标是估计出设定密度 $f(y|\mathbf{x},\boldsymbol{\theta})$ 的参数。为了简单起见,假定总体分层概率 H_s 已知。科斯利特(Cosslett,1981a)首先通过设 $\mathbf{x}$ 是以概率 w_i 出现 $\mathbf{x}_i$ 的一种离散形式,然后对联合似然求关于 $\boldsymbol{\theta}$ 与 w_i 的极大值,$i=1,\cdots,N$,得到具有内生分层的 MLE。为了得出仅仅包括 $(q+S-1)$ 个参数 $\boldsymbol{\theta}$ 与函数 $\lambda_s(\boldsymbol{\theta})$ 的一种集中似然函数,其一阶条件会失效。其次,对这个集中似然函数求关于 $\boldsymbol{\theta}$ 与 λ_s 的极大值,从而得出的估计值与求关于 $\boldsymbol{\theta}$ 与 $\lambda_s(\boldsymbol{\theta})$ 的极大值相同。最后,由于将 λ_s 处理成参数是有效的,故同样的方法能用于连续回归元的情况。维数 q 加上无穷维未知密度 $h(\mathbf{x})$ 的问题被简化成 $q+S-1$ 维。

GMM 估计

科斯利特(Cosslett,1981a)的著名结果很难实施。

英伯斯(Imbens, 1992)曾推导出较简单的具有内生分层的 GMM 估计量,该估计量的效果与科斯利特的 MLE 一样。针对通过多项式抽样、标准分层抽样或可变概率抽样所得到的分层样本,英伯斯和兰开斯特(Imbens and Lancaster, 1996)给出了这种估计量的十分一般的框架与阐述。其联合密度再次是式(24.16)的 $g^{s}(s,y,\mathbf{x}|\boldsymbol{\theta})$,并且允许样本层概率 H_s 可能是未知的。GMM 分析建立在下面三种方程与一个最终约束的基础上。具体地讲,这三种方程分别是:H_s 得分的 $S-1$ 个方程;基于给定 s 与 $\mathbf{x}$ 时 y 的条件似然函数的 $\boldsymbol{\theta}$ 的 q 个方程;关于总体层概率 $Q_s(\boldsymbol{\theta})$ 约束的 $S-1$ 个方程。最终约束是,当 $Q_s(\boldsymbol{\theta})$ 存在线性约束时,可不必用该约束,例如,如果层是互斥的且覆盖样本空间,就会出现这种情况。

24.4.5 *加权估计*

当样本层与总体层即由式(24.14)定义的 H_s 与 $Q_s(\boldsymbol{\theta})$ 都是已知时，内生分层很容易处理，尽管估计量不是完全有效的。我们在考察更一般的估计量之前，以 ML 估计开始。

加权 ML 估计

曼斯基和莱尔曼(Manski and Lerman, 1977)曾经提出加权极大似然(**weighted maximum likelihood**，记为 WML)估计量。该估计量对：

$$Q_{\mathrm{WML}}(\boldsymbol{\theta}) = \sum_i \frac{Q_i}{H_i}\ln f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}) \tag{24.17}$$

求极大值，其中 $H_i = H_s$，而 $Q_i = Q_s$，当第 i 个观测值位于 s 层时。

曼斯基和莱尔曼(Manski and Lerman, 1977)称此估计量为加权外生抽样估计量(**weighted exogenous sampling estimator**，记为 WESML)，因为式(24.17)用权数 H_i/Q_i 乘以在外生抽样的条件似然中的通常项 $\ln f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta})$。不过，称谓 WESML 能够引起混淆，因为该问题是一个内生性问题，即恰好可以证明，对通常外生估计量进行适当加权就得到一致估计量。

沿着类似线索，目标函数 $Q_{\mathrm{WML}}(\boldsymbol{\theta})$ 正式地讲不是似然函数，因为式(24.16)并不蕴含给定 $\mathbf{x}$ 与 s 时，y 的样本条件密度是由 $f^s(y \mid \mathbf{x}, \boldsymbol{\theta}) = f(y \mid \mathbf{x}, \boldsymbol{\theta})^{Q_s/H_s}$ 给出。不过，WML 估计量是一致的。WML 估计量是一阶条件

$$\sum_i \frac{Q_i}{H_i}\frac{\partial \ln f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0} \tag{24.18}$$

的解。当求和式中的项具有零期望值时，期望针对关于式(24.16)中的抽样密度 $g^s(s, y, \mathbf{x} \mid \boldsymbol{\theta})$ 而取，则这个估计量是一致的。现在，在通常正则条件下，即就总体而言，设定密度满足 $\mathrm{E}[\partial \ln f(y \mid \mathbf{x}, \boldsymbol{\theta})/\partial \boldsymbol{\theta}] = \mathbf{0}$，得到：

$$\begin{aligned}
&\mathrm{E}_s\left[\frac{Q_s}{H_s}\frac{\partial \ln f(y \mid \mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] \qquad (24.19)\\
&= \iint \frac{Q_s}{H_s}\frac{\partial \ln f(y \mid \mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\frac{H_s}{Q_s(\boldsymbol{\theta})} f(y \mid \mathbf{x}, \boldsymbol{\theta}) h(\mathbf{x}) dy d\mathbf{x}\\
&= \iint \frac{\partial \ln f(y \mid \mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f(y \mid \mathbf{x}, \boldsymbol{\theta}) h(\mathbf{x}) dy d\mathbf{x}\\
&= \int \mathrm{E}\left[\frac{\partial \ln f(y \mid \mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] h(\mathbf{x}) d\mathbf{x}\\
&= \mathbf{0}
\end{aligned}$$

因此，在有内生分层情况下，WML 估计量一致的。

对于式(24.17)中的目标函数 $Q_{\mathrm{WML}}(\boldsymbol{\theta})$ 来说，信息矩阵等式不成立，所以我们需要使用 $\hat{\boldsymbol{\theta}}_{\mathrm{WML}}$ 的渐近方差的三明治形式 $N^{-1}\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$，其中：

$$\mathbf{A}(\boldsymbol{\theta}_0) = \operatorname{plim}\frac{1}{N}\sum_{i=1}^{N}\frac{Q_i}{H_i}\left.\frac{\partial^2 \ln f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}_0} \tag{24.20}$$

与

$$\mathbf{B}(\boldsymbol{\theta}_0) = \text{plim}\frac{1}{N}\sum_{i=1}^{N}\left(\frac{Q_i}{H_i}\right)^2 \frac{\partial \ln f(y_i|\mathbf{x}_i,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ln f(y_i|\mathbf{x}_i,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}_0} \quad (24.21)$$

与科斯利特或莫伯斯的 ML 估计量相比，这个估计量的有效性会差一些，但是它实施起来相对简单。当然，它要假定有层概率知识。

加权 m 估计

加权 ML 估计量能应用到除条件 ML 之外的估计量。例如，豪斯曼和怀斯(Hausman and Wise, 1979)考察了类似的关于最小二乘回归的加权估计。

因而，假定满足 SRS，我们对 $\sum_i q(y_i|\mathbf{x}_i,\boldsymbol{\theta})$ 求极小值，得到一阶条件 $\sum_i \partial q(y_i|\mathbf{x}_i,\boldsymbol{\theta})/\partial\boldsymbol{\theta}=\mathbf{0}$，并且对于总体，假定：

$$\mathrm{E}[\partial q(y|\mathbf{x},\boldsymbol{\theta})/\partial\boldsymbol{\theta}]=0$$

这是一致性的必要条件。那么，抽样反而是内生分层的，如同 24.2 节一样，同时样本层 H_s 与总体层 Q_s 均为已知，则 $\boldsymbol{\theta}$ 由加权 m 估计量得到一致估计 $\hat{\boldsymbol{\theta}}_{\mathrm{W}}$，该估计量 $\hat{\boldsymbol{\theta}}_{\mathrm{W}}$ 极小化：

$$Q_{\mathrm{W}}(\boldsymbol{\theta}) = \sum \frac{Q_i}{H_i} q(y_i|\mathbf{x}_i,\boldsymbol{\theta}) \quad (24.22)$$

对于 WML 估计量，其一致性证明由式(24.18)与式(24.19)可得，而其方差矩阵的形式为 $N^{-1}\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$，其中，$\mathbf{A}$ 与 $\mathbf{B}$ 已由式(24.20)与式(24.21)给出，只是唯一变动是由 $\partial q(y_i|\mathbf{x}_i,\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ 代替 $\partial \ln f(y_i|\mathbf{x}_i,\boldsymbol{\theta})/\partial\boldsymbol{\theta}$。伍德里奇(Wooldridge, 2001)给出了一种正式证明。

类似地，在内生分层下，对于基于 q 个总体矩条件 $\mathrm{E}[\mathbf{h}(y,\mathbf{x},\boldsymbol{\theta})]=\mathbf{0}$ 的估计来说，使用加权估计方程估计量(**weighted estimating equations estimator**)，该估计量是

$$\sum_i \frac{Q_i}{H_i}\mathbf{h}(y_i,\mathbf{x}_i,\boldsymbol{\theta}) = \mathbf{0}$$

的解。加权 MLE 结果用到了，$h(y_i|\mathbf{x}_i,\boldsymbol{\theta})$ 用代替 $\partial \ln f(y_i|\mathbf{x}_i,\boldsymbol{\theta})$。

注意到，加权 Q_i/H_i 与 24.3.2 节在简单外生分层抽样下关于普查参数估计所提出的那些一样。不过，其动机截然不同。本节假定，条件矩得以正确设定，所以就外生分层抽样而言，实施非加权估计会是一致的且有效的。如果分层是内生的，就必须进行加权。

24.5 聚 集

关于加权及分层的 24.3 节与 24.4 节内容涵盖了调查设计的一些方法，此类调查设计会导致样本分布不同于总体分布。抽样观测值独立性的假设要继续保持。

实际上，调查数据经常是相依的。这可能是为了减少调查成本而使用聚集样本，比如对同一街区访问成千上万个住户。在此类情况下，数据因为有共同的不可

观测特定群项，在同一类中可能出现相关。可是，甚至对于 SRS 来说，也会出现这种相依性。例如，可能认为，对同一个州的所有住户来说，存在不可观测的共同影响。

存在几种控制群内的不可观测相依性的不同方法。如果群内不可观测因素与回归元不相关，那么仅有回归参数的方差需要加以调整。相反，若群内不可观测因素与回归元相关，则回归元参数估计都是非一致的，从而需要其他可供选择的合适估计量。依照是存在许多小群还是几种大群，方法还可能有所不同，所以分析起来极为复杂。另外，复杂调查的新困难，比如加权及分层，则推迟到 24.6 节讨论。

下面将阐述随机群效应与固定群效应之间的重要区别，其记号与模型类似于面板数据分析。下面各个小节阐述各种不同估计量。

24.5.1 *特定群效应模型*

关注内容在于给定数据$(y_i, \mathbf{x}_i)$时线性回归模型的估计，$i=1,\cdots,N$，其中，i 表示第 i 个样本观测值，比如住户。

考虑的内容是，总体回归模型的某些方面随群 c 而变化，$c=1,\cdots,C$。假定在整个样本中的第 i 个住户是第 c 个抽样群的第 j 个住户。对于聚集数据来说，一种相当一般的模型是：

$$y_{jc}=\mathbf{x}_{jc}'\boldsymbol{\beta}_c+\mu_{jc}, \quad j=1,\cdots,N_c, \ c=1,\cdots,C \tag{24.23}$$

其中，$\mathrm{Cov}[u_{jc}, u_{kc}]\neq 0$，尽管对于 $c\neq d$，$\mathrm{Cov}[u_{jc}, u_{kd}]=0$。该模型通过下述方式并入了群相依性：其方式是既包括回归参数随不同群而变化，又包括误差在某一群内是相关的。

这里，我们把焦点放在一种情况，即特定群效应模型（**cluster-specific effects model**，记为 CSEM）：

$$y_{jc}=\mathbf{x}_{jc}'\boldsymbol{\beta}+\alpha_c+\varepsilon_{jc} \tag{24.24}$$

此处，回归截距 α_c 恰好随不同群而变化，而其斜率系数被假定成对不同群来说是一个常值。在最简单模型中，ε_{jc} 被假定成同方差的：

$$\varepsilon_{jc}\sim[0,\sigma_\varepsilon^2] \tag{24.25}$$

为了允许异方差性与群内相关，对该假设加以放松。总的来讲，对 α_c 做出不同假设，就会导致两个截然不同的模型，现在就阐述它们。

特定群随机效应

就特定群随机效应（CSRE）模型而言，式（24.24）的截距 α_c 是纯随机的，其分布不依赖于任何观测因素。在最简单情况下，假定：

$$\alpha_c\sim[0,\sigma_\alpha^2] \tag{24.26}$$

这个模型非常类似于面板数据的随机效应模型。该模型刚好是 y_{jc} 关于 $\mathbf{x}_{jc}$ 的一个线性回归，其新的复杂情况是，误差项 $\alpha_c+\varepsilon_{jc}$ 因同一群观测因素而成为相关的。

OLS 估计是一致的，却是无效的。重要的是，误差相关必须引起对 OLS 估计

量的通常标准误差加以调整。GLS 估计则是更有效的。

已知关于 ε_{jc} 与 α_c 的假设(24.25)与(24.26),有 $V[\alpha_c+\varepsilon_{jc}]=\sigma_\alpha^2+\sigma_\varepsilon^2$,并且 $Cov[\alpha_c+\varepsilon_{jc},\alpha_c+\varepsilon_{kc}]=\sigma_\alpha^2$,对于 $k\neq j$。我们定义如下的群内相关系数(**intraclass correlation coefficient**):

$$\rho=Cor[\alpha_c+\varepsilon_{jc},\alpha_c+\varepsilon_{jc}]=\frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\varepsilon^2} \tag{24.27}$$

这是$(\sigma_\alpha^2,\sigma_\varepsilon^2)$与$(\sigma^2,\rho)$之间的一种对应,其中,$\rho$ 已由式(24.2)定义,而 $\sigma^2=\sigma_\alpha^2+\sigma_\varepsilon^2$。CSRE 模型等价于具有常值群内相关系数的模型。对该模型也可给出一种贝叶斯解释,将每个观测值看成拥有自己的截距 α_{jc},而 α_{jc} 是来自多元分布的一个采样,并且令人感兴趣的可交换性准则如下:α_{jc} 中的下标只是一个标记符号而已,并无实质结果。就一切情况而论,聚集具有引起群内误差项之间正相关的期望效应。

特定群固定效应

就特定群固定效应(**cluster-specific random effects**,记为 CSFE)模型而言,式(24.23)的截距 α_c 是随机不可观测的,至于 CSRE 模型,却可能与回归元相关。为了识别,$\mathbf{x}_{jc}$ 不再包括截距项。

这个 CSFE 模型非常类似于面板数据的固定效应模型。该模型有条件均值 $E[y_{jc}|\mathbf{x}_{jc},\alpha_c]=\mathbf{x}_{jc}'\boldsymbol{\beta}+\alpha_c$。当省略变量 α_c 与 $\mathbf{x}_{jc}$ 相关时,来自 y_{jc} 对 $\mathbf{x}_{jc}$ 回归的 OLS 估计量 $\boldsymbol{\beta}$ 仅仅关于是非一致的。

对 $\boldsymbol{\beta}$ 进行一致估计要求对 α 有一致估计,假如群很大,则可能出现此情况。相反,当群很小,个体 α_c 就需要通过差分变换加以剔除。

与面板数据分析的比较

很明显,设置背景与术语都密切地极相似于第 21～23 章曾经阐述的静态面板数据分析。同时,存在着某些背离面板数据分析的地方。

在面板数据情况下,对个体比如住户加以分析,对该个体单元观测不止一次,然而在聚集情况下,分析的个体单元仅仅观测一次。在面板数据记号里,当面板是一个短面板时;第 1 个下标表示聚集单元。然而,在聚集记号里,jc 的第 2 个下标则是聚集单元。在面板情况下,我们关注平衡面板,而当 N_c 随不同群变化时,聚集数据时常是非平衡的。

面板数据方面的微观经济计量学聚焦于短面板。这类似于拥有每群仅有几个观测值且众多群的情况。于是,当 N_c 是小的且 $C\to\infty$,我们称为小群(**small cluster**)。另外,出现大群(**large cluster**)也很平常,即当 $N_c\to\infty$ 且 C 是小的那种群。对于含有大群的 CSFE 模型来说,少数几个参数 α_c 要去掉,并且不会出现非主要参数问题。

与面板数据不同,适当的聚集单元可能并不总是清晰可见的。例如,就 CPS 数据而言,聚集可能被看成在州内引起的或在层内引起的;在 PSU 内引起的或在 USU 内引起的。这个问题推迟到 24.6 节讨论。群内相关被认为是对在更高汇总水平上聚集减少。若聚集在州水平出现,则该群是大的,然而若将聚集看成在 USU 水平出现的,则群是小的。此外,一种可能情况是,数据集合并不包括必需的集群信息,比如层或 USU 作为观测值。

动态面板数据而非表态面板数据模型的类似形式是，模型中的 y_{jc} 不仅依赖于 $\mathbf{x}_{jc}$，而且依赖于 $\mathbf{x}_{kc}$，$k \neq j$。对于聚集数据来说，通常足以设定一种同伴效应模型(**peer-effects model**)，该模型更直接地包含群平均值$\bar{\mathbf{x}}_c$，因为群内的观测值次序通常不起作用。

概述集群方面的三个普遍估计量是 24.5.2～24.5.4 节阐述的 OLS、GLS 和组内估计量。这些估计量的性质已由表 24.2 节概括出来，性质会随真实模型而变化。特别重要的是，若真实模型是特定群固定效应的，则 OLS 与 RE 估计量都是非一致的，然而组内估计量却会得到一致估计值，只是仅对那种群内变化的回归元系数。其次，即使回归元是一致的，为了控制集群与下面将要详述的可能异方差性，经常需要对通常标准误差加以调整。

表 24.2　各种不同聚集模型估计量的性质

节	估计量	群模型	一致性
24.5.2	OLS	随机效应	是
		固定效应	不是
24.5.3	随机效应的 GLS	随机效应	是
		固定效应	不是
24.5.4	群内固定效应	随机效应	是
		固定效应	是

24.5.2　*OLS 估计量*

我们考察 OLS 回归：

$$y_{jc}=\mathbf{x}_{jc}'\boldsymbol{\beta}+u_{jc} \tag{24.28}$$

普通 OLS 是非一致的，因为有省略变量偏倚，当真实模型是 CSFE 模型(即 $u_{jc}=\alpha_c+\varepsilon_{jc}$)，其固定效应 α_c 与 $\mathbf{x}_{jc}$ 相关。那么，就不应使用 OLS 估计量，而应运用 24.5.4 节的 CSFE 估计量。

与之相比，CSRE 模型的 OLS 则是一致的，其 α_c 是随机效应并与 $\mathbf{x}_{jc}$ 不相关。更一般地讲，倘若 u_{jc} 与 $\mathbf{x}_{jc}$ 不相关，则在比 CSRE 模型更为丰富的 u_{jc} 模型条件下，OLS 是一致的。我们考察此情况下的 OLS 估计量，关注已知群内误差项 u_{jc} 的相关性时求正确标准误差。

记号

对式(24.28)群内观测值加以叠放，得出：

$$\mathbf{y}_c=\mathbf{X}_c\boldsymbol{\beta}+\mathbf{u}_c \tag{24.29}$$

其中，$\mathbf{y}_c$ 与 $\mathbf{u}_c$ 均表示 $N_c\times1$ 维向量，$\mathbf{X}_c$ 表示 $N_c\times K$ 阶矩阵。进一步地，对不同群叠放，得到：

$$\mathbf{y}=\mathbf{x}\boldsymbol{\beta}+\mathbf{u} \tag{24.30}$$

其中，$\mathbf{y}$ 与 $\mathbf{u}$ 均表示 $N\times1$ 维向量，$\mathbf{X}$ 表示 $N\times K$ 阶矩阵，$N=\sum_c N_c$。

CSRE 模型的三种表示会产生描述模型(24.28)中 OLS 估计量的三种等价方式：

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{OLS}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \Big(\sum_{c=1}^{C}\mathbf{X}_c'\mathbf{X}_c\Big)^{-1}\sum_{c=1}^{C}\mathbf{X}_c'\mathbf{y}_c \\ &= \Big(\sum_{c=1}^{C}\sum_{j=1}^{N_c}\mathbf{x}_{jc}\mathbf{x}_{jc}'\Big)^{-1}\sum_{c=1}^{C}\sum_{j=1}^{N_c}\mathbf{x}_{jc}y_{jc}\end{aligned} \tag{24.31}$$

当已知误差对不同群是独立假设时，这些表述中的第二种形式尤其有用。从而，如同前面面板情况，OLS 估计量具有极限分布：

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{\text{OLS}}-\boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}[\mathbf{0},\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}] \tag{24.32}$$

其中：

$$\begin{aligned}\mathbf{A} &= \text{plim}\, N^{-1}\sum_{c=1}^{C}\mathbf{X}_c'\mathbf{X}_c \\ \mathbf{B} &= \text{plim}\, N^{-1}\sum_{c=1}^{C}\mathbf{X}_c'\mathbf{u}_c\mathbf{u}_c'\mathbf{X}_c\end{aligned} \tag{24.33}$$

这里用到了 $\mathbf{u}_c$ 关于 c 是独立的。对 $\mathbf{u}_c$ 做出不同假设会得出 $\mathbf{B}$ 的各种不同估计值。

OLS 群稳健的标准误差

如果一些群是小的，那么存在许多群，当用 $\hat{\mathbf{u}}_c=\mathbf{y}_c-\mathbf{X}_c\hat{\boldsymbol{\beta}}$ 代替 $\mathbf{u}_c$ 时，式(24.33)的 $\mathbf{B}$ 就能被一致地估计出。由此可得，$\hat{\boldsymbol{\beta}}_{\text{OLS}}$渐近服从正态分布，其群稳健方差矩阵：

$$\hat{\text{V}}[\hat{\boldsymbol{\beta}}_{\text{OLS}}] = \Big(\sum_{c=1}^{C}\mathbf{X}_c'\mathbf{X}_c\Big)^{-1}\sum_{c=1}^{C}\mathbf{X}_c'\hat{\mathbf{u}}_c\hat{\mathbf{u}}_c'\mathbf{X}_c\Big(\sum_{c=1}^{C}\mathbf{X}_c'\mathbf{X}_c\Big)^{-1} \tag{24.34}$$

这个公式没有对异方差性与群内相关性施加约束，从而 $\text{V}[\mathbf{u}_c]$是无约束的，因此 $\text{V}[u_{jc}]$与 $\text{Cov}[u_{jc},u_{kc}]$也无约束。不过，假定 N_c 是小的且 $C\to\infty$。统计软件包经常给出自由度修正。典型地讲，人们用

$$dfc=\frac{N-1}{N-K}\times\frac{C}{C-1}$$

乘以式(24.34)的估计，这既是对 $\boldsymbol{\beta}$ 的估计进行校正，也是对实际应用中群的数目成为有限的校正。

为了理解式(24.34)是如何起作用的，将回归元处理成固定的，并注意到：

$$\begin{aligned}\mathbf{B} &= \lim N^{-1}\sum_{c=1}^{C}\mathbf{X}_c'\text{E}[\mathbf{u}_c\mathbf{u}_c']\mathbf{X}_c \\ &= \lim N^{-1}\sum_{c=1}^{C}\sum_{j=1}^{N_c}\sum_{k=1}^{N_c}\text{E}[u_{jc}u_{kc}']\mathbf{x}_{jc}\mathbf{x}_{kc}'\end{aligned}$$

于是，利用估计值：

$$\begin{aligned}\hat{\mathbf{B}} &= N^{-1}\sum_{c=1}^{C}\mathbf{X}_c'\hat{\mathbf{u}}_c\hat{\mathbf{u}}_c'\mathbf{X}_c \\ &= N^{-1}\sum_{c=1}^{C}\sum_{j=1}^{N_c}\sum_{k=1}^{N_c}\hat{u}_{jc}\hat{u}_{kc}\mathbf{x}_{jc}\mathbf{x}_{kc}'\end{aligned}$$

得到式(24.34)。例如，考察用 $\bar{y}$ 对 $\text{E}[y]$的估计。这是满足 $\mathbf{x}_{jc}=1$、$\hat{\boldsymbol{\beta}}_{\text{OLS}}=\bar{y}$ 并且

$\hat{u}_{jc}=y_{jc}-\bar{y}$ 的回归(24.28)。从而,由式(24.34)得出 $\mathrm{V}[\bar{y}]=N^{-2}\sum_c(\sum_j(y_{jc}-\bar{y}))^2$。与 $N^{-1}\sum_c\sum_j(y_{jc}-\bar{y})^2$ 的估计相比,此处额外假定了群内的独立性。

CSRE 模型的 OLS 标准误差

群稳健估计(24.34)需要有众多群。如果对模型误差 u_{jc} 的方差与协方差做出某些假设,那么就可使用一种可供选择的估计,该估计应用于仅有几个群的情况。利用这些可供选择的估计量,可以得到有关聚集对估计方差影响的分解结果。

特别地,假定由式(24.24)与式(24.26)给出的 CSRE 模型是合适的。那么,误差 $u_{jc}=\alpha_c+\epsilon_{jc}$ 对不同 c 来说是独立的,并且在群内有:

$$\mathrm{Cov}[u_{jc},u_{kc}]=\begin{cases}\sigma^2, & j=k,\\ \rho\sigma^2, & j\neq k,\end{cases}$$

其中,群内相关系数 ρ 已由式(24.27)定义。由此可得:

$$\boldsymbol{\Sigma}_c=\mathrm{V}[\mathbf{u}_c]=\sigma^2[(1-\rho)\mathbf{I}_c+\rho\mathbf{e}_c\mathbf{e}_c'] \tag{24.35}$$

其中,$\mathbf{I}_c$ 表示一个 $N_c\times N_c$ 阶单位矩阵,$\mathbf{e}_c$ 表示一个元素为 1 的 $N_c\times 1$ 维向量。

已知式(24.35)的 $\boldsymbol{\Sigma}_c$,由一般性结果(24.32)与(24.33)可得:

$$\mathrm{V}[\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}]=\left(\sum_{c=1}^{C}\mathbf{X}_c'\mathbf{X}_c\right)^{-1}\sum_{c=1}^{C}\sigma^2\mathbf{X}_c'[(1-\rho)\mathbf{I}_c+\rho\mathbf{e}_c\mathbf{e}_c']\mathbf{X}_c\left(\sum_{c=1}^{C}\mathbf{X}_c'\mathbf{X}_c\right)^{-1} \tag{24.36}$$

倘若群内相关系数是常值,这个方差矩阵估计量不论是在小群还是大群情况下均为一个常值。关于 σ^2 与 ρ 的明显估计量分别是:

$$\hat{\sigma}^2=\frac{1}{N-K-1}\sum_{c=1}^{C}\sum_{j=1}^{N_c}\hat{u}_{jc}^2$$

与

$$\hat{\rho}=\frac{1}{\sum_c N_c(N_c-1)}\frac{1}{\hat{\sigma}^2}\sum_{c=1}^{C}\sum_{j=1}^{N_c}\sum_{k\neq j}^{N_c}\hat{u}_{jc}\hat{u}_{kc}$$

ρ 的估计涉及许多群内对,而且一致估计可通过利用群内对的子集来获得。写成 $\sum_c N_c(N_c-1)$ 的对都是有用的,虽然每一个唯一的群内对实际上被成对地记录成 $\hat{u}_{jc}\hat{u}_{kc}$ 与 $\hat{u}_{kc}\hat{u}_{jc}$ 而出现在求和中。

如果群是大的,就允许群内相关随不同群而变化。于是,式(24.35)与式(24.36)能借助于用 σ_c^2 与 ρ_c 分别代替 σ^2 与 ρ 而得到修正。这里,σ_c^2 与 ρ_c 可用

$$\hat{\sigma}_c^2=\frac{1}{N_c-K-1}\sum_{j=1}^{N_c}\hat{u}_{jc}^2$$

与

$$\hat{\rho}_c=\frac{1}{N_c(N_c-1)}\frac{1}{\hat{\sigma}_c^2}\sum_{j=1}^{N_c}\sum_{k\neq j}^{N_c}\hat{u}_{jc}\hat{u}_{kc}$$

得到一致估计。

通常 OLS 标准误差的偏倚

当数据出现聚集时,从直观上看,OLS 估计量的通常方差估计量公式是:

$$V^{\text{Formula}}[\hat{\boldsymbol{\beta}}_{\text{OLS}}]=\sigma^2\left(\sum_{c=1}^{C}\mathbf{X}_c'\mathbf{X}_c\right)^{-1}$$

上式低估了 OLS 估计值的真实方差矩阵，假定群内有正相关，由于群内每增加一个额外观测值将提供少于独立信息增加一个额外值时的信息量。我们阐述误差过程为 CSRE 模型的特有形式时的这种偏倚。考察一种 CSRE 模型，在每一个群内其回归元都是相同的，因此 $\mathbf{x}_{jc}=\mathbf{x}_c$，且 $\mathbf{X}_c=\mathbf{e}_c\mathbf{x}_c'$。于是，通过利用 $\mathbf{e}_c'\mathbf{e}_c=N_c$，式(24.36)变成：

$$V[\hat{\boldsymbol{\beta}}_{\text{OLS}}]=\left(\sum_{c=1}^{C}N_c\mathbf{x}_c\mathbf{x}_c'\right)^{-1}\sum_{c=1}^{C}N_c\sigma^2[1+\rho(N_c-1)]\mathbf{x}_c\mathbf{x}_c'\left(\sum_{c=1}^{C}N_c\mathbf{x}_c\mathbf{x}_c'\right)^{-1}$$

此结果是由克勒克(Kloek，1981)和莫尔顿(Moulton，1986)提出的。

现在，对平衡群进行专门研究，并定义 M 是平均群容量，所以 $M=N_c=N/C$ 为常值。从而，方差估计可简化成：

$$V[\hat{\boldsymbol{\beta}}_{\text{OLS}}]=[1+\rho(M-1)]\times\sigma^2\left(M\sum_{c=1}^{C}N_c\mathbf{x}_c\mathbf{x}_c'\right)^{-1}$$

而其方差公式简化成 $\sigma^2(M\sum_c\mathbf{x}_c\mathbf{x}_c')^{-1}$。由此可得，真实方差是下面这个数值：

$$\tau=[1+\rho(M-1)]$$

乘以通常 OLS 方差矩阵估计。即使 ρ 是小的，校正因子也会相当大。例如，若平均群容量为 $M=101$ 个观测值，则应用 $\sqrt{1+100\rho}$ 乘以通常 OLS 标准误差。对每个群内所假定的独立性也会得出 σ^2 的一个有偏估计，但这是拥有三阶的重要性形式。在平衡群情况下，克勒克已经证明，$\mathrm{E}[\sum_c\sum_j\hat{u}_{cj}^2]=\sigma^2[N-K(1+\rho(m-1))]$，所以我们应该用$[N-K(1+\rho(m-1))]^{-1}$而不是$[N-K]^{-1}$加以正规化。

在实际应用中，某些回归在一个群内可能为常值，而另一些回归元则可能变化。那么，在回归拥有截距及纯量回归元(即 $\mathbf{x}_{jc}'\boldsymbol{\beta}=\beta_1+\beta_2x_{jc}$，)的情况下，斯科特和霍尔特(Scott and Holt，1982)已经证明，关于截距的通常 OLS 方差公式应该用 $1+\rho(M-1)$去乘，如同前面所述，只是对于斜率系数来说，则应该用较小因子 $1+\hat{\rho}_x\rho(M-1)$去乘，其中，$\hat{\rho}_x$ 被看成是 x_{jc} 的群内相关系数的估计值。在横截面应用中，$\hat{\rho}_x$ 相对较小，因此，主要问题出在群不变回归元的标准误差上。

莫尔顿(Moulton，1986)在一个应用中阐述，利用错误的 OLS 方差公式时标准误差中的偏倚是相当大的。他运用横截面 CPS 数据估计了对数工资方程，其中，聚集出现在州水平上。就他的应用而言，$N=18.946$ 而 $C=49$。针对他的数据，估计群内相关系数为 $\hat{\rho}=0.032$，看起来似乎很小。不过，群是大的，并且我们忽略是非平衡的，运用上面平均群容量 $M=387$ 的公式作为指南，则有 $\hat{\tau}=[1+\hat{\rho}(M-1)]=13.3$。就州不变回归元而言，可以预测真实 OLS 标准误差是 $\sqrt{13.3}=3.7$ 乘以通常报告的标准误差，它是一个极大的偏倚。(看待该种情形的一种方法是，对于州不变回归元系数的 OLS 估计来说，18 946 个整群观测值具有相同的精度，因为有 18 946/13.3=1 425 个独立观测值。)就个体变化回归元而言，其偏倚将会更小一些，例如当 $\hat{\rho}_x=0.10$ 时，$[1+\hat{\rho}_x\hat{\rho}(M-1)]=2.23$。莫尔顿没有报告结果，因为个体变化回归包括了回归元。对于州不变回归元，诸如某个州的就业增长率这样的变

量，OLS 的群校正标准误差一般位于 3 倍错误标准误差公式与 4 倍错误标准误差公式之间。

一个教训是，对于群不变性回归元的 OLS 系数，其默认 OLS 标准误差中存在很大的向下偏倚。就个体变化回归元而言，也存在偏倚，只是它更小而已。带有整群数据的应用经常包含群不变回归元。有效的统计推断需要获得控制聚集的标准误差工。

24.5.3 *特定群随机效应估计量*

如果随机效应模型合适，那么 GLS 估计量一般比上一节的 OLS 估计量更为有效。已知对于不同群具有独立性，模型(24.29)的 GLS 估计量是：

$$\hat{\boldsymbol{\beta}}_{\mathrm{GLS,RE}} = \Big(\sum_{c=1}^{C}\mathbf{X}_c'\boldsymbol{\Sigma}_c^{-1}\mathbf{X}_c\Big)^{-1}\sum_{c=1}^{C}\mathbf{X}_c'\boldsymbol{\Sigma}_c^{-1}\mathbf{y}_c \tag{24.37}$$

其中，$\boldsymbol{\Sigma}_c=\mathrm{V}[\mathbf{u}_c]$。可行的 GLS 估计量是用 $\boldsymbol{\Sigma}_c$ 的一致估计量 $\hat{\boldsymbol{\Sigma}}_c$ 代替 $\boldsymbol{\Sigma}_c$，一旦假定正确设定模型(24.29)以及误差方差矩阵 $\boldsymbol{\Sigma}_c$，则有：

$$\mathrm{V}[\hat{\boldsymbol{\beta}}_{\mathrm{GLS,RE}}] = \Big(\sum_{c=1}^{C}\mathbf{X}_c'\boldsymbol{\Sigma}_c^{-1}\mathbf{X}_c\Big)^{-1}$$

对于 CSRE 模型，式(24.35)给出的 $\boldsymbol{\Sigma}_c$ 能用 $\hat{\boldsymbol{\Sigma}}_c$ 一致地估计出，其中，σ^2 与 ρ 都要用式(24.36)后面给出的一致估计值加以代替。如同面板数据的随机效应模型一样，可行 GLS 估计量渐近地等价于另一假设，即 α_c 与 ε_{jc} 均服从正态分布下的 MLE。

CSRE 模型吸引人的地方是，GLS 估计量(24.37)能直接借助于变换回归

$$y_{jc}-\theta_c\bar{y}_c=(\mathbf{x}_{jc}-\theta_c\bar{\mathbf{x}}_c)'\boldsymbol{\beta}+(\varepsilon_{jc}-\theta_c\bar{\varepsilon}_c) \tag{24.38}$$

的 OLS 估计而得以实施。其中：

$$\theta_c=1-\frac{\sqrt{1-\rho}}{\sqrt{1+\rho(N_c-1)}}=1-\frac{\sigma_\varepsilon}{\sqrt{\sigma_\varepsilon^2+N_c\sigma_\alpha^2}} \tag{24.39}$$

本节稍后将证明该结果。为了实施式(24.37)，我们用 θ_c 的一致估计值 $\hat{\theta}_c$ 代替 θ_c。如同面板数据模型，可以证明，能够运用来自这个回归的通常 OLS 标准误差，当模型(24.24)中的误差 ε_{jc} 是同方差时。

当假定式(24.24)与式(24.26)成立时，GLS 估计量至少与 OLS 同样有效。在所有回归元都是整群不变的特殊情况下，因为 GLS 与 OLS 相一致，所以有效性没有提高[克勒克(Kloek, 1981)]。更一般地，斯科特和霍尔特(Scott and Holt, 1982)给出了与 GLS 相比，OLS 有效性损失的一个相当保守的上界：

$$\frac{\mathrm{V}[\mathbf{c}'\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}]}{\mathrm{V}[\mathbf{c}'\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}]}\geqslant 1-\left(1+\frac{4(1-\rho)[1+\rho(N_0-1)]}{N_0^2\rho^2}\right)^{-1}$$

其中，对于任意向量 $\mathbf{c}$，$N_0=\max\{N_c\}$ 是最大群的样本量。该界随 N_0 与 ρ 而增大，甚至对于 $N_0=1\,000$ 与 $\rho=0.10$ 来说，与 GLS 相比，OLS 有效性至多损失 22%。

针对 GLS 而言，已知，这些有效性提高很少，一种更为普遍的方式是关注含有正确标准误差的 OLS 估计，除非由于 CSFE 模型合适，OLS 就是非一致的。聚集的重要影响是，与那些没有出现聚集情况相比，OLS 的有效性表现得更差，这一点可以从 24.5.2 节对 OLS 估计量标准误差的计算讨论中明显地看出来。

当整群是大的，就对 CSRE 模型加以放松，以使误差方差与群内相关随不同群而变化。于是，对式(24.35)的 $\mathbf{\Sigma}_c$，我们分别用式(24.36)给出的 σ^2 与 ρ 的一致估计值 σ_c^2 与 ρ_c 代替它们。

当整群是小的，类似于 OLS 的式(24.34)，可以得出一种稳健的标准误差，该误差并没有迫使误差相关在群内成为常值。于是：

$$\hat{\mathrm{V}}[\hat{\boldsymbol{\beta}}_{\mathrm{GLS,RE}}]=\Big[\sum_{c=1}^{C}\mathbf{X}_c'\mathbf{\Sigma}_c^{-1}\mathbf{X}_c\Big]^{-1}\sum_{c=1}^{C}X_c'\hat{\mathbf{\Sigma}}_c^{-1/2}\hat{\mathbf{u}}_c\hat{\mathbf{u}}_c'\hat{\mathbf{\Sigma}}_c^{-1/2}\mathbf{X}_c\Big[\sum_{c=1}^{C}\mathbf{X}_c'\hat{\mathbf{\Sigma}}_c^{-1}\mathbf{X}_c\Big]^{-1}$$

其中，$\hat{\mathbf{u}}_c=\mathbf{y}_c-\mathbf{X}_c\hat{\boldsymbol{\beta}}_{\mathrm{GLS,RE}}$。该估计要求 N_c 很小且 $C\to\infty$，并假定误差在不同群之间具有独立性。

将 GLS 实施成变换模型的 OLS

为了推导出式(24.38)，注意到，对于式(24.35)定义的 $\mathbf{\Sigma}_c$，有：

$$\begin{aligned}\mathbf{\Sigma}_c^{-1}&=[\sigma^2[(1-\rho)\mathbf{I}_c+\rho\mathbf{e}_c\mathbf{e}_c']]^{-1}\\&=\frac{1}{\sigma^2(1-\rho)}[\mathbf{I}_c-(\rho/\tau_c)\mathbf{e}_c\mathbf{e}_c']^{-1}\end{aligned}$$

其中，$\tau_c=1+\rho(N_c-1)$，从而：

$$\mathbf{\Sigma}_c^{-1}=\frac{1}{\sigma\sqrt{1-\rho}}[\mathbf{I}_c-(\theta_c/N_c)\mathbf{e}_c\mathbf{e}_c']$$

这里用到了一般性结果：若 $\mathbf{e}$ 是 $M\times1$ 维向量，其元素都为 1，则：

$$\begin{aligned}&[\mathbf{I}+a\,\mathbf{ee}']^{-1}=\mathbf{I}-[a/(1+aM)]\mathbf{ee}'\\&[\mathbf{I}+a\,\mathbf{ee}']^{1/2}=\mathbf{I}-M^{-1}(1-\sqrt{1+aM})\mathbf{ee}'\end{aligned}$$

现在，式(24.37)中的 $\mathbf{X}_c'\mathbf{\Sigma}_c^{-1}\mathbf{X}_c=(\mathbf{\Sigma}_c^{-1/2}\mathbf{X}_c)'\mathbf{\Sigma}_c^{-1/2}\mathbf{X}_c$，其中：

$$\begin{aligned}\mathbf{\Sigma}_c^{-1/2}\mathbf{X}_c&=[\mathbf{I}_c-(\theta_c/N_c)\mathbf{e}_c\mathbf{e}_c']\mathbf{X}_c\\&=\mathbf{X}_c-\theta_c\mathbf{e}_c\bar{\mathbf{x}}_c'\end{aligned}$$

而 $\bar{\mathbf{x}}_c'=N_c^{-1}\sum_j\mathbf{x}_{jc}$，我们可忽略纯量倍数 $1/\sigma\sqrt{1-\rho}$，因为当我们类似地考察 $\mathbf{X}_c'\mathbf{\Sigma}_c^{-1}\mathbf{y}_c$ 时，它将被消掉。从而，得到变换回归模型(24.38)。

24.5.4 *特定群固定效应估计量*

CSFE 模型的基本思想简单朴素：设群效应通过截距项引入条件均值函数之中。该模型是：

$$y_{jc}=\alpha_c+\mathbf{x}_{jc}'\boldsymbol{\beta}+\varepsilon_{jc},\quad j=1,\cdots,N_c,\ c=1,\cdots,C \tag{24.40}$$

现在，$\boldsymbol{\beta}$ 与 α_c 都是待估参数，$c=1,\cdots,C$。

在 CSFE 模型中，所有整群不变的回归元必须被去掉，因为它们并不能从 α_c 中独立地识别出。例如，聚集出现在州水平上，且固定效应模型合适，那么就不能识别州不变回归元比如州平均失业率的效应。假如人们希望估计州不变回归元的系数，此时反而需要运用 OLS 或 CSRE 估计量。不过，人们应首先使用类似于面板数据第 21 章阐述的豪斯曼检验，验证 CSRE 模型的强假设：α_c 与回归元不相关的有效性。

我们考察在假设：

$$\varepsilon_{jc} \sim [0, \sigma_{jc}^2]$$

下的统计推断。这允许有异方差性的未知形式，但假定包含特定群固定效应 α_c，这样做足以控制群内的任何误差相关。

这背离了面板数据分析；涉及误差方面的时间序列相关，甚至在通常包含特定个体效应之后，产生更为丰富模型的情形。不过，如果人们愿意，就能通过类似 24.5.2 节中的那些方法，另外调整群内相关的标准误差估计量。

估计 CSFE 模型的一个主要新困难是，小群因存在众多截距 α_c 而无法估计出来。

群虚拟变量模型

首先，我们考虑大群，其中，群数相对于总样本量而言相对较小。于是，截距 α_c 能通过直接进入每个群的虚拟变量并用 OLS 估计出来。设观测值 i 表示第 c 个群的第 j 个住户。那么，可将式(24.40)写成一种群虚拟变量模型(**cluster dummy variables model**)：

$$y_i = \sum_{c=1}^{C} \alpha_c d_{ci} + \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \cdots, N \tag{24.41}$$

其中，d_{ci} 表示指示变量，当第 i 个观测值属于 c 群时，$d_{ci}=1$，否则 $d_{ci}=0$。因而，C 个群指示变量比如州虚拟变量，都被包括进来，并且为了避免出现虚拟变量困境，不应包括截距项。

对这种模型进行 OLS 估计，不仅会得到 $\alpha_1, \cdots, \alpha_C$ 的一致估计值，也会得到 $\boldsymbol{\beta}$ 的一致估计值，一旦假定群数 C 固定而 $N \to \infty$ 时。人们能使用通常艾克—怀特估计，获得对给定异方差误差而言是稳健的标准误差。

群内估计量

当有许多小群时，我们就不可以再通过 OLS 估计模型(24.40)。首先，由于当群数 $C \to \infty$ 时，参数个数$(C+K) \to \infty$，所以从计算形式上看，OLS 估计行不通。其次，也是更重要的，因为参数数量随样本量趋于无穷大，除非 $N_c \to \infty$，否则 OLS 估计量是非一致的。

通常，关注内容在于式(24.40)中的参数 $\boldsymbol{\beta}$，将 $\alpha_1, \cdots, \alpha_C$ 看成非主要参数或冗余参数[1](**nuisance parameter**)。那么，一种方便的方法是，通过对最初数据作变换，清除固定效应。每一个观测值$(y_{jc}, \mathbf{x}_{jc})$都要用其与群平均的离差代替，即

[1] 又称为多余参数。——译者注

$(y_{jc}-\bar{y}_c, \mathbf{x}_{jc}-\bar{\mathbf{x}}_c)$，$i=1,\cdots,N_c$，$c=1,\cdots,C$，其中，$\bar{y}_c=N_c^{-1}\sum_j y_{jc}$，而 $\bar{\mathbf{x}}_c=N_c^{-1}\sum_j \mathbf{x}_{jc}$ 是特定群平均值。于是，关于 y_{jc} 的模型(24.40)蕴含：

$$y_{jc}-\bar{y}_c=(\mathbf{x}_{jc}-\bar{\mathbf{x}}_c)'\boldsymbol{\beta}+\varepsilon_{jc}-\bar{\varepsilon}_c \tag{24.42}$$

对变换回归(24.42)运用 OLS，得到 $\boldsymbol{\beta}$ 的一致估计。当 CSFE 系数也是关注内容时，能通过 $\hat{\alpha}_c=\bar{y}_c-\bar{\mathbf{x}}_c'\boldsymbol{\beta}$ 加以估计，尽管这种估计对于小 N_c 来说不是一致的。

与第 21 章相比，可以证明，这是类似于面板数据的组内估计量。至于面板数据，来自对式(24.42)进行 OLS 估计所得到的 $\boldsymbol{\beta}$ 估计值，与来自对群虚拟变量模型(24.41)进行 OLS 估计所得到的 $\boldsymbol{\beta}$ 估计值是一致的。

类似于线性面板模型，也可以提出一种群间估计量(**between estimator**)。在此情况下，$\bar{y}_c$ 对 $\bar{\mathbf{x}}_c$ 进行回归，由式(24.37)知，CSRE 模型的 GLS 估计量涉及准差分形式的回归，其中，在进行差分之前要用 θ_c 乘以群均值[由式(24.39)定义]。可以证明，GLS 估计量是群内估计量与群间估计量的一种线性组合。从而，当 $\theta_c\to 1$ 时，它接近于大 N_c 的群内估计量。注意到，CSRE 模型中的群内估计量是一致的。

当用修正均值观测值进行回归时，解释标准误差就要小心谨慎。因为这种回归的自由度数目是$(N-K-C)$，而不是$(N-K)$。如果软件忽略了这种调整，那么由软件得到的残差方差应该用扩张因子$(N-K)/(N-K-C)$去乘，而其标准误差则应该用扩张因子的平方根去乘。

24.5.5 *对群效应的诊断检验*

在线性回归中，在误差服从正态条件下，对特定群固定效应进行检验刚好是式(24.40)的线性约束假设 $H_0: \alpha_1=\alpha_2=\cdots=\alpha_C=0$ 的标准 F 检验。这直接需要对含有特定群虚拟变量回归与没有特定群虚拟变量回归的两种 R^2 统计量加以比较。

在 CSRE 模型中，群效应的检验是零假设 $H_0:\sigma_\alpha^2=0$ 对 $H_1: \sigma_\alpha^2>0$ 的单侧检验。一种等价检验也可以表述成：利用式(24.27)定义的 $H_0: \rho=0$ vs $H_1: \rho>0$ 的检验。该检验的单侧 LM 检验统计量由莫尔顿(Moulton, 1987)提出，即：

$$\mathrm{LM}=\frac{\sum_c(N_c\bar{u}_c)^2-\sum_c\sum_i\hat{u}_{ic}^2}{\hat{\sigma}^2[2(\sum_c N_c^2-N)]^{1/2}} \tag{24.43}$$

其中，$\hat{\sigma}^2=\sum_c\sum_i\hat{u}_{ic}^2/N$，$\hat{u}_{ic}$ 表示实施 y 对 $\mathbf{x}$ 的混合回归而得到的最小二乘残差，$\bar{u}_c$ 表示群 c 的平均残差。

24.5.6 *非线性模型聚集*

经济计量学文献中，带有整群数据的非线性模型并没有引起人们更多注意。不过，在生物统计学里出版了大量文章，特别关注于二值结果模型[彭德格斯特等人(Pendergast et al., 1996)]，而且考虑了其他一些模型，比如泊松回归、生存数据的某些模型。特别是，分层(多水平)建模框架也被广泛用于二值结果模型上。

这里，我们继续探索整群数据与面板数据之间的平行内容。如同线性情况，数据$(y_i,\mathbf{x}_i)$，$i=1,\cdots,N$ 被写成标出下标形式的$(y_{jc},\mathbf{x}_{jc})$，$j=1,\cdots,N_c$，$c=1,\cdots,C$，假定对不同 c 有独立性，却允许在群 c 内观测值具有相关性。

针对聚集的m估计

考察一种非线性估计方程估计量，该估计量是

$$\sum_{c=1}^{C}\sum_{j=1}^{N_c}\mathbf{h}(y_{jc},\mathbf{x}_{jc},\boldsymbol{\theta})=\mathbf{0} \tag{24.44}$$

的解。这些方程经常通过对目标函数$\sum_c\sum_j q(y_{jc},\mathbf{x}_{jc},\boldsymbol{\theta})$求极大值或求极小值而获得，在此情况下，$\mathbf{h}(y_{jc},\mathbf{x}_{jc},\boldsymbol{\theta})=\partial q(y_{jc},\mathbf{x}_{jc},\boldsymbol{\theta})/\partial\boldsymbol{\theta}$。例如，建立在边缘密度$\mathbf{h}(y_{jc},\mathbf{x}_{jc},\boldsymbol{\theta})=\partial\ln f(y_{jc}|\mathbf{x}_{jc},\boldsymbol{\theta})/\partial\boldsymbol{\theta}$乘积基础上的拟MLE。

我们假定数据是整群的，因而$\mathrm{Cov}[\mathbf{h}_{jc},\mathbf{h}_{kc}]\neq\mathbf{0}$。不过，保留一致性必要条件：$\mathrm{E}[\mathbf{h}(y_{jc},\mathbf{x}_{jc},\boldsymbol{\theta})]=\mathbf{0}$的假设，这排除下面将阐述的特定群固定效应模型。

很容易对OLS估计量(24.34)的群稳健方差加以校正，以便适应当前情况，即用$\partial\mathbf{h}_{jc}/\partial\boldsymbol{\theta}'$代替$\mathbf{x}_{jc}\mathbf{x}_{jc}'$，并且用$\mathbf{h}_{jc}(\hat{\boldsymbol{\theta}})$代替$\mathbf{x}_{jc}\hat{u}_{jc}$。于是，$\hat{\boldsymbol{\theta}}$渐近服从正态分布，其群稳健方差矩阵为：

$$\hat{\mathrm{V}}[\hat{\boldsymbol{\theta}}]=\left(\sum_{c=1}^{C}\sum_{j=1}^{N_c}\frac{\partial\mathbf{h}_{jc}'}{\partial\boldsymbol{\theta}}\bigg|_{\hat{\theta}}\right)^{-1}\sum_{c=1}^{C}\sum_{j=1}^{N_c}\sum_{k=1}^{N_c}\mathbf{h}_{jc}(\hat{\boldsymbol{\theta}})\mathbf{h}_{kc}(\hat{\boldsymbol{\theta}})'\left(\sum_{c=1}^{C}\sum_{j=1}^{N_c}\frac{\partial\mathbf{h}_{jc}}{\partial\boldsymbol{\theta}'}\bigg|_{\hat{\theta}}\right)^{-1} \tag{24.45}$$

有的计算软件将此作为许多非线性参数模型的一个标准选项。

一个重要例子是，建立群内边缘密度乘积而不是联合密度基础上的拟ML估计。具体地讲，已知对群内c的不同j具有相关性，我们应极大化对数似然：

$$\ln\mathrm{L}(\boldsymbol{\theta})=\sum_{c=1}^{C}\ln f(y_{1c},\cdots,y_{N_c c},\ \mathbf{x}_{1c},\cdots,\mathbf{x}_{N_c c},\boldsymbol{\theta})$$

不过，以联合密度开始研究很困难，或者很难获得联合密度，因为对于许多一元密度来说，多元密度存在的范围有限。然而，我们可以极大化：

$$\begin{aligned}Q(\boldsymbol{\theta})&=\sum_{c=1}^{C}\ln[f(y_{1c},\mathbf{x}_{1c},\boldsymbol{\theta})\times\cdots\times f(y_{N_c},\mathbf{x}_{N_c},\boldsymbol{\theta})]\\&=\sum_{c=1}^{C}\sum_{j=1}^{N_c}\ln f(y_{jc},\ \mathbf{x}_{jc},\boldsymbol{\theta})\end{aligned}$$

这不再是真实似然函数，除非y_{jc}对不同j是独立的，因此信息矩阵不再能应用。运用$\mathbf{h}_{jc}(\boldsymbol{\theta})=\partial\ln f(y_{jc},\mathbf{x}_{jc},\boldsymbol{\theta})/\partial\boldsymbol{\theta}$且$\partial\mathbf{h}_{jc}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'=\partial^2\ln f(y_{jc},\mathbf{x}_{jc},\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$，前面公式得到应用。

这意味着，在每个群内，我们不能利用每个观测值的似然得分，如同存在独立观测值的情形；相反，我们要用整群元素上的似然得分之和代替它。

非线性特定群随机效应

非线性模型的特定群效应的相当一般设置是，考察极大化或极小化下式的估计量：

$$Q(\boldsymbol{\beta},\alpha_1,\cdots,\alpha_C)=\sum_{c=1}^{C}\sum_{j=1}^{N_c}q(y_{jc},\mathbf{x}_{jc},\boldsymbol{\beta},\alpha_c) \tag{24.46}$$

其中，群效应经由纯量参数α_c而引入，$c=1,\cdots,C$。一种简单的随机效应模型假定α_c是iid的，并具有参数$\boldsymbol{\delta}$。关于α_c取期望，得出目标函数：

$$Q(\boldsymbol{\beta},\delta)=\sum_{c=1}^{C}\int\sum_{j=1}^{N_c}q(y_{jc},\mathbf{x}_{jc},\boldsymbol{\beta},\alpha_c)f(\alpha_c|\boldsymbol{\delta})d\alpha_c$$

特别地，当此和式积分不存在闭形式表达式时，估计起来极为复杂。

经常容易得到关于一个观测值的期望，$E_{\alpha_c}[q(y_{jc},\mathbf{x}_{jc},\boldsymbol{\beta},\alpha_c)]=q^*(y_{jc},\mathbf{x}_{jc},\boldsymbol{\beta},\boldsymbol{\delta})$。于是，一种较简单的估计量是忽略聚集，并求 $Q^*(\boldsymbol{\beta},\delta)=\sum_c\sum_j q^*(y_{jc},\mathbf{x}_{jc},\boldsymbol{\beta},\boldsymbol{\delta})$ 的极小值，该估计值是一致的，尽管对于聚集来说，需要利用式(24.45)的标准误差加以调整。

例如，针对计数数据，我们发展一种面板泊松—伽玛混合模型的整群数据形式。不过，忽略了聚集的泊松拟 MLE 还是能得到应用的，因为它是一致的，尽管针对聚集情况，标准误差需要加以调整。

因此，即使人们发展了非线性模型的随机效应形式，但一种适宜的方法是，经常通过忽略聚集后对参数进行估计，然后对于聚集情况修正估计量的标准误差。除潜在提高有效性以外，很少有理由去估计整群随机效应模型。

非线性特定群固定效应

特定群固定效应模型的非线性变形是对

$$Q(\boldsymbol{\beta},\alpha_1,\cdots,\alpha_C)=\sum_{c=1}^{C}\sum_{j=1}^{N_c}q(y_{jc},\mathbf{x}_{jc},\boldsymbol{\beta},\alpha_c)$$

求极大值或极小值，如同式(23.34)[1]一样，只是现在参数 $\alpha_1,\cdots,\alpha_C$ 是待估的，而不是积分去掉。

对于大整群，即 C 小且 $N_c\to\infty$，我们仅仅对 $Q(\boldsymbol{\beta},\alpha_1,\cdots,\alpha_C)$ 求关于 $\boldsymbol{\beta}$ 与 $\alpha_1,\cdots,\alpha_C$ 的最优值。一旦假定 $\alpha_1,\cdots,\alpha_C$ 完全控制任意聚集，推断就能建立在通常由 iid 假设获得的标准误差基础上。这是特定群虚拟变量模型(24.41)的非线性类似形式。

对于小整群，即 N_c 小且 $C\to\infty$，就会出现太多的非主要参数 $\alpha_1,\cdots,\alpha_C$ 问题。与线性模型不同，一般地讲，不可能去掉参数 $\alpha_1,\cdots,\alpha_C$[霍尔和塞弗尼(Hall and Severini, 1998)]。不过，由第 23 章面板数据，我们看到，在一些情况下去掉参数 $\alpha_1,\cdots,\alpha_C$ 是可能的。

例如，含有群固定效应的二值 logit 模型设定：

$$\Pr[y_{jc}=1]=\frac{1}{1+\exp(-\alpha_c-\mathbf{x}_{jc}'\boldsymbol{\beta})}\tag{24.47}$$

其中，为了识别，$\mathbf{x}_{jc}$ 不能包括截距或群不变回归元。固定效应 α_c 能利用条件 MLE 加以去掉，这里的条件 MLE 是以群内响应之和 $\sum_{j=1}^{N_c}y_{jc}=N_c\bar{y}_c$ 为条件的。第 c 个群的联合条件概率是：

$$\Pr[y_{1c},\cdots,y_{N_cc}|N_c\bar{y}_c]=\frac{\exp(\boldsymbol{\beta}\sum_{j=1}^{N_c}\mathbf{x}_{jc}y_{jc})}{\sum_{d\in\tilde{B}_c}\exp(\boldsymbol{\beta}\sum_{j=1}^{N_c}\mathbf{x}_{jc}d_{jc})}\times\frac{\Gamma[\sum_{j=1}^{N_c}y_{jc}+1]\Gamma[N_c-\sum_{j=1}^{N_c}y_{jc}+1]}{\Gamma(N_c+1)}\tag{24.48}$$

[1] 原著中这里为式(24.34)，应为式(23.34)。——译者注

其中，$\widetilde{B}_c=\{(d_{1c},\cdots,d_{N_c c})\mid d_{nc}=0$ 或 1，且 $\sum_j d_{jc}=\sum_j y_{jc}\}$。条件似然是关于所有群项比如这些项的一个乘积，而容量为 1 的群则被该似然函数排斥在外。右边第二项不依赖于未知参数，从而不会影响到似然函数的极大化，因此，当考虑极大化时可以忽略它。该似然函数不便于求极大值，因为集合 $\widetilde{B}_c$ 涉及从群 c 的总结果 $(N_{1c}+N_{0c})$ 选取 N_c 个结果 $y_{jc}=1$ 的许多方式。不过幸运的是，大量流行的计算机软件都提供了用于估计这种模型的条件 logit 选项。所有未知参数的协方差矩阵，可通过对数似然海赛矩阵的逆得以估计出来。

举另一个例子，考察泊松固定效应群模型，它设定：

$$y_{jc}\sim\mathcal{P}[\mu_{jc}=\alpha_c\exp(\mathbf{x}_{jc}'\boldsymbol{\beta})],\quad c=1,\cdots,C$$

其中，$\mathcal{P}[\cdot]$表示泊松分布，而 $\mathbf{x}_{jc}$ 不包括截距与任何群不变回归元。这是一般的泊松模型，只是通常条件均值 $\exp(\mathbf{x}_{jc}'\boldsymbol{\beta})$用特定群固定效应 α_c 去乘。对于这种特殊模型，一系列方法包括条件 ML 与中心化 ML 都会去掉参数 α_c。借助于求解估计方程：

$$\sum_{c=1}^{C}\sum_{j=1}^{N_c}\mathbf{x}_{jc}\left(y_{jc}-\frac{\bar{y}_c}{\bar{\lambda}_c}\lambda_{jc}\right)=\mathbf{0}$$

而得到 $\boldsymbol{\beta}$ 参数的一致估计。其中，$\lambda_{jc}=\exp(\mathbf{x}_{jc}'\boldsymbol{\beta})$，$\bar{y}_c=N_c^{-1}\sum_j y_{jc}$，$\bar{\lambda}_c=N_c^{-1}\sum_j\lambda_{jc}$，它们都是群均值。对于更详细内容，参见第 23.7 节在面板数据情况下对此类问题的讨论。

24.5.7 *整群数据其他方法*

聚集的一个基本特征是，对于不同观测值出现相依性。一个有关的专题是空间相关(**spatial correlation**)[参见安塞林(Anselin, 2001)、李明宰(Lee, 2004)]，其中，观测单元是一个地区，比如州，相互邻接地区的观测值可能是相关的。

为了考察斜率系数与截距，就要推广随机效应方法。这是下一节分层线性模型(**hierarchical linear models**)要阐述的。对于非线性模型，问题类似于第 23 章所述的面板数据内容。

在聚集产生群内相关却不影响估计量的一致性背景下，运用自助法来获得群稳健标准误差。从直观上看，人们应对群 c 采用放回再抽样，在此情况下，我们要求满足 $C\to\infty$的小整群。针对第 b 次自助，我们以放回方式采样 C 个群，并利用这 C 个再抽样群中的全部 j 个住户去估计$\hat{\boldsymbol{\theta}}_b$，$\hat{\boldsymbol{\theta}}_b$ 是式(24.44)的解。于是，人们通过将通常样本方差公式应用于 $\hat{\boldsymbol{\theta}}_1,\cdots,\hat{\boldsymbol{\theta}}_B$，估计出 $\mathrm{V}[\hat{\boldsymbol{\theta}}]$，其中，$B$ 表示自助复制次数。注意到，再抽样是针对整群而不是住户实施的，因为，假定整数是 iid 的，而存在群内相依性。

24.6 分层线性模型

24.5 节将随机效应模型的群效应作用限制成为回归截距。更一般的随机效应模型，还会允许群差异体现在斜率系数上。体现在回归系数的子集的群间差异

与可观测的群特征相联系。由于此类模型包括了几个设定层次，所以称为分层模型（**hierarchical models**）。

许多应用统计学科中关于整群数据的标准框架是分层线性模型，又称多水平线性模型[1]（**multilevel linear models**）、随机系数模型、方差成分模型、混合线性模型或混合效应模型（**mixed effects models**）。模型的这种分类带来了设定方面的更多信息。我们以对分组个体群模型阐述开始讨论。该种模型适合于短面板数据情况，即对每一个体来说，重复测量数据出现聚集。

24.6.1 *模型结构*

分层或多水平模型是能用于带有嵌套结构的数据的一类模型。一些例子包括，某一地区诸如州或郡的个体，或者是某一个组织单位比如学校或社会的个体，或者某个家庭个体，比如可以利用的有双胞胎数据。面板数据也是一个例子，这里将对同一个个体的重复测量值解释成嵌入个体的观测值。

我们以线性模型：

$$y_i = \mathbf{x}'_{ij}\boldsymbol{\beta}_j + u_{ij} \tag{24.49}$$

开始讨论，其中，新项目是设 K 个回归元系数 $\boldsymbol{\beta}$ 随组（或群）j 而变化。一个具体例子是，考虑学校内学生数据。于是，y_{ij} 是一个结果测量值，比如第 j 个学校的第 i 个学生的测验分数，而回归元比如学生民族变动的边际效应会随学校不同而变化。注意到，我们使用的标准分层线性模型（HLM）记号，与 24.5 节所用的那些记号相比恰好相反，其中，y_{cj} 表示第 c 个学校的第 j 位学生的测验分数。

两水平分层线性模型，将第一水平模型（24.49）中的系数设定成由一个随机项与第二水平变量的线性函数来决定，这里的第二水平变量是学校特征。以纯量参数 β_{kj} 开始阐述，即 $K\times 1$ 维向量参数 $\boldsymbol{\beta}_j$ 的第 k 个分量。从而，将 β_{kj} 建模成依赖于学校特征向量 $\mathbf{w}_k$ 并满足：

$$\beta_{kj} = \mathbf{w}'_{kj}\boldsymbol{\gamma}_k + v_{kj}, \quad k=1,\cdots,K \tag{24.50}$$

的模型，这里，$\mathbf{w}_k$ 对第 j 个学校取值 $\mathbf{w}_{kj}$，$\mathbf{w}_{kj}$ 的第一个分量通常为常数。若对 $\boldsymbol{\beta}$ 的所有 K 个成分叠放，则有：

$$\begin{bmatrix}\beta_{1j}\\ \vdots \\ \beta_{Kj}\end{bmatrix} = \begin{bmatrix}\mathbf{w}'_{1j} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \ddots & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{w}'_{Kj}\end{bmatrix}\begin{bmatrix}\gamma_1\\ \vdots \\ \gamma_K\end{bmatrix} + \begin{bmatrix}v_{1j}\\ \vdots \\ v_{Kj}\end{bmatrix}$$

或用明显矩阵记号写成：

$$\boldsymbol{\beta}_j = \mathbf{W}_j\boldsymbol{\gamma} + \mathbf{v}_j \tag{24.51}$$

模型（24.50）是一种灵活形式，并且嵌入许多模型作为特殊情况。这些特殊情况包括含有随机截距与随机斜率的模型，但该框架还允许回归系数随第二水平可观测而变化。模型范围极为广泛，正如下面指出的那样。

当 $\beta_{kj}=\gamma_k$ 时，第 k 个第一水平系数称为固定系数，在此情况下，系数不随第二水平回归元或不可观测因素而变化。如果所有第一水平系数都是固定的，那么模

[1] 又称为多层线性模型。——译者注

型(24.49)简化成 $y_{ij}=\mathbf{x}'_{ij}\boldsymbol{\gamma}+u_{ij}$,在此情况下通过估计回归就合适。注意到,与面板背景下经济计量学家使用的固定效应项相比,固定系数项具有截然不同的意义。

第 k 个第一水平系数称为非随机变化系数,如果 $\beta_{kj}=\gamma_k$。那么,该系数是学校特征的一个线性函数。假如所有第一水平系数都是固定的,只是截距是非随机变化的,则模型(24.49)简化成 $y_{ij}=\mathbf{x}'_{ij}\boldsymbol{\gamma}+u_{ij}$,在此情况下,通过 OLS 估计回归就合适。注意到,与面板背景下经济计量学家所使用的固定效应项相比,固定系数项具有截然不同的意义。

第 k 个第一水平系数成为非随机变化系数,如果 $\beta_{kj}=\mathbf{w}'_{kj}\boldsymbol{\gamma}_k$,那么,该系数是学校特征的一个线性函数。假如所有第一水平系数都是固定的,只是截距是非随机变化得,则模型(24.49)简化成 $y_{ij}=\mathbf{x}'_{ij}\boldsymbol{\beta}+\mathbf{w}'_{1j}\boldsymbol{\gamma}_1+u_{ij}$,它是关于个体特性与学校特征结果的一个标准 OLS 回归。

第 k 个第一水平系数被称为随机变化系数,如果 $\beta_{kj}=\gamma_k+v_{kj}$。那么,该系数是纯随机的,并且不随学校特征而变化。倘若所有第一水平系数都是随机变化的,因而 $\boldsymbol{\beta}_j=\boldsymbol{\gamma}+\mathbf{v}_j$,则模型是方差分量模型或随机系数模型。当所有第一水平系数都是固定的,只是截距是随机变化的,模型(24.49)简化成 $y_{ij}=\mathbf{x}'_{ij}\boldsymbol{\beta}+v_{1j}+u_{ij}$,它是一个随机截距模型。

在实际应用中,某些第一水平系数既是非随机变化的又是随机变化的,如同一般情况下的式(24.49)。假如仅有第一水平截距遵从一般模型(24.49),而其他所有第一水平系数都为固定的,则模型(24.49)简化成 $y_{ij}=\mathbf{x}'_{ij}\boldsymbol{\beta}+\mathbf{w}'_{1j}\boldsymbol{\gamma}_1+v_{1j}+u_{ij}$。这是通常的混合回归模型,误差有两个分量,因此,误差对在同一个学校内的不同个体来说是相关的。

HLM 框架能被推广到更多水平上。例如,个体学生(下标 i)可以被嵌入学校(下标 j),学校被嵌入某个地区(下标 k)。于是,三个水平 HLM 在第一水平上将学生结果设定成 $y_{ijk}=\mathbf{z}'_{ijk}\boldsymbol{\pi}_{jk}+e_{ijk}$,其中,参数 $\boldsymbol{\pi}_{jk}=\mathbf{X}_{jk}\boldsymbol{\beta}_k+\mathbf{u}_{jk}$,同样地有 $\boldsymbol{\beta}_k=\mathbf{W}_j\boldsymbol{\gamma}+\mathbf{w}_k$。

HLM 可被重新写成一种混合线性模型,因为将式(24.50)代入式(24.49)得到:

$$y_{ij}=(\mathbf{x}'_{ij}\mathbf{W}_j)\boldsymbol{\gamma}+\mathbf{x}'_{ij}\mathbf{v}_j+u_{ij} \tag{24.52}$$

目标是估计回归参数 $\boldsymbol{\gamma}$、误差 u_{ij} 与 $\mathbf{v}_j$ 的方差与协方差。由于假定误差与回归元独立,所以对式(24.52)的混合 OLS 估计得出 $\boldsymbol{\gamma}$ 的一致参数回归。HLM 方法运用了更有效的估计量,这些估计量利用对误差 u_{ij} 与 $\mathbf{v}_j$ 的方差及协方差所做的假设。

在最简单情况下,被假定成是 iid 的,服从 $\mathcal{N}[0,\sigma^2]$,而 $\mathbf{v}_j$ 被假定成 iid 的,服从 $\mathcal{N}[\mathbf{0},\boldsymbol{\Gamma}]$。于是,模型能重新写成:

$$y_{ij}\sim\mathcal{N}[\mathbf{x}'_{ij}\boldsymbol{\beta}_j,\sigma^2]$$
$$\boldsymbol{\beta}_j\sim\mathcal{N}[\mathbf{W}_j\boldsymbol{\gamma},\boldsymbol{\Gamma}]$$

在贝叶斯背景下,林德利和史密斯(Lindley and Smith,1972)对这个模型做出了早期的研究,其中,$\boldsymbol{\gamma}$ 被称为超参数,在更一般模型中,超参数本身同样依赖于更高水平的超参数。参数 $\boldsymbol{\gamma}$、σ^2、$\boldsymbol{\Gamma}$ 可通过极大似然法或贝叶斯方法估计出来。作为一种选择方式,能够使用 ML 方法,这在本质上与 21.7 节所述的混合线性面板数据的

那种情况一样。布雷克和劳登布什(Bryk and Raudenbush, 1992, 2002)给出了一个完整研究。

24.6.2 面板数据的 HLM

HLM 文献将短面板解释成对个体的重复测量。于是,个体就成为两个水平 HLM 的第二水平,而在上一节里,个体则是第一水平。

模型(24.28)变成:

$$y_{ti}=\mathbf{x}_{ti}'\boldsymbol{\beta}_i+u_{ti} \tag{24.53}$$

比如,这里 y_{ti} 表示第 i 个学生在时间 t 的一个测量结果,并且诸如所学特定科目的回归元变动的边际效应随学生不同而变化。纯量参数 β_{ki},即 $K\times 1$ 维向量参数 $\boldsymbol{\beta}_i$ 的第 k 个元素,被建模成依赖于个体特征 $\mathbf{w}_k$ 向量,这里对 $\mathbf{w}_k$ 第 i 个个体来说取值为 $\mathbf{w}_{ki}$,满足:

$$\beta_{ki}=\mathbf{w}_{ki}'\gamma_k+v_{ki} \tag{24.54}$$

特定个体效应模型是下面情况的一种特殊情形:所有的第一水平系数是固定的,因而 $\beta_{ki}=\gamma_k$,只是截距项 β_{1i} 会随个体不同而变化(第二水平分组)。

当截距 β_{1i} 不存在任何模型而直接估计 β_{1i} 时,就是特殊个体固定效应模型。这是满足 $\beta_{1i}=\mathbf{w}_{1i}'\gamma_1$ 的非随机变化系数的极端情况,其中,$\mathbf{w}_{1i}$ 表示 $N\times 1$ 维指示变量向量,当 $i=l$ 时,第 l 个分量等于 1,否则等于 0,因此,$\beta_{1i}=\gamma_{1i}$。HLM 框架并不是设计成适应经济计量学家所称谓的固定效应模型。

当截距 β_{1i} 是一个随机变化系数时,因而 $\beta_{1i}=\gamma_1+v_{1i}$,即特定个体随机效应模型。很明显,人们能设定出更具一般性的随机效应模型,β_{ki} 同样依赖于回归元。

正如已经注意到的,HLM 是一种混合线性模型。对于面板数据情况,式(24.52)的类似形式是:

$$y_{ti}=(\mathbf{x}_{ti}'\mathbf{W}_i)\boldsymbol{\gamma}+\mathbf{x}_{ti}'\mathbf{v}_j+u_{ti}$$

对第 21 章的随机效应模型加以专门研究,得出 $y_{ti}=\mathbf{x}_{ti}'\boldsymbol{\gamma}+v_j+u_{ti}$。

HLM 框架在面板数据方面的标准应用是增长模型,其中,结果 y_{ti} 是个体智力或身高,它是年龄的函数,并且允许年龄的边际效应随个体不同而变化。这里除允许截距随个体不同而变化之外,还允许斜率系数随个体不同而变化。

24.7 聚集例子:越南保健支出

在本节,我们关注存在聚集时的估计,因为这在微观经济计量学研究中是调查数据方面最为普通的复杂情况。可以运用 24.5 节的方法。

不论是线性回归模型还是非线性回归模型,都建立在来自世界银行关于越南生活标准调查(VLSS)1997~1998 年个体水平及住户水平数据的基础上。一系列专项的详细信息调查收集源自大致 6 000 个住户的 27 700 个个体,住房分布在大约 194 个社区。下文将"社区"处理成群或组,并假定观测结果与所处社区是相关的。住户样本平均群容量大约为 26,最大群容量是 39,而最小群容量为 1。为了阐明线性群模型与非线性群模型,对三个结果加以建模。

第一，我们考察住户保健年度总支出的线性回归模型（LNEXP12M），对于拥有正支出的住户，作为住户总支出（HHEXP）（对数）函数，它控制几个标准社会人口统计变量，即保健支出的“恩格尔曲线”类型。关注内容是住户总支出的系数，它是保健需求关于住户收入强性的估计值。

第二，我们运用个体响应信息，估计保健类型的整群计数模型，以此解释汇总私人保健支出的高比例。在对这些结果进行建模时，我们控制个体的最近健康状况、家庭收入、健康保险状态以及各种人口统计变量，比如年龄、婚姻状况、户主的受教育水平。健康状况被限制在调查期间维持的 ILLNESS 或 INJURY、生病期间还有受限活动天数。关注的重要系数再次是收入与保险状态变量的系数。

表 24.3 汇集了这些例子所用的变量定义与概括统计量。

表 24.3　越南保健例子所用数据描述

住户数据	定义	均值	标准差
LNEXP12M	12 个月的全部住户保健支出	6.31	1.59
AGE	户主年龄	48.01	13.77
SEX	当户主为女性则取 1，否则 0	0.27	0.44
HHSIZE	全部住户的人口数	4.73	1.96
URBAN	当住户为城市的则取 1，否则 0	0.29	0.45
EDUC	户主受教育年限	7.09	4.41
HHEXP	全部住户名义支出（1998 年越南盾）	15 273	13 020
个体数据			
PHARVIS	直接去药店的次数	0.51	1.31
LNMEDEXP(>0)	那些有正支出住户的 log（总医疗支出）	2.14	1.08
AGE	年龄	29.7	9.67
SEX	若回答者为男性，则等于 1	0.51	0.49
MARRIED	若为已婚者，则等于 1	0.40	0.49
EDUC	获得的毕业文凭水平	3.38	1.94
ILLNESS	在过去 12 个月里没有得病数	0.62	0.90
INJURY	若在调查期间受伤，则等于 1	0.62	0.90
ILLDAYS	得病天数	2.80	5.45
ACTDAYS	受限制活动天数	0.06	1.11
INSURANCE	若回答者有医疗保险，则等于 1	0.16	0.37
MEDEXP(>0)	以正医疗支出为条件的医疗支出	21.04	208
MEDEXP	医疗支出（1998 越南盾）	6.13	112.75

这两种情况的核心问题如下：聚集对弹性估计影响是多少？当运用各种不同的假设、模型以及估计量时，弹性及其影响会怎样变化？

24.7.1　*结果与讨论*

表 24.4 给出了 OLS 回归、HC t 比率、固定效应以及随机效应公式表述结果。与来自运用不考虑聚集的异方差——致方差估计量所得到的标准误差结果相比，相对变化并不太。不过，当使用聚集稳健方差估计量（24.34）时，其标准误差出现了相当大的变动。支出弹性的 t 比率从 16.01 下降到 12.68。所有 t 比率都变得更小了，而且变量 SEX 与 HHSIZE 的下降到 1.96 以下。正如人们所料，这些结果显示，忽略群间相关性引发了 OLS t 比率增大。

表 24.4 越南保健:正支出的 FE 模型与 RE 模型

	OLS				FGLS		FE		RE(GLS)	
变量[a]	系数	OLS	异方差$\|t\|$	群$\|t\|$	系数	$\|t\|$	系数	$\|t\|$	系数	$\|t\|$
LNHHEXP	0.670	16.01	15.76	12.68	0.620	14.14	0.603	11.61	0.626	13.29
AGE	0.010	6.39	6.36	5.46	0.011	6.96	0.011	6.93	0.011	6.85
SEX	0.097	1.88	1.88	1.64	0.108	2.13	0.112	2.17	0.106	2.09
HHSIZE	0.028	2.19	2.15	1.89	0.014	1.06	0.010	0.76	0.015	1.17
FARM	0.134	2.73	2.72	2.22	0.088	1.58	0.069	1.14	0.092	1.69
EDUC	−0.090	7.36	7.07	6.03	−0.061	4.73	−0.051	3.76	−0.063	4.92
CONS	−0.510	1.34	1.34	1.09	−0.051	0.30	−0.051	0.08	−0.166	0.40
R^2	0.088									
R_W^2									0.051	
R_B^2									0.288	
ρ					0.12					
$\frac{\sigma_\alpha^2}{(\sigma_\alpha^2+\sigma^2)}$									0.093	
$F(193, 4806)$							3.49			
$\chi^2(1)$									432.75	
豪斯曼 $\chi^2(6)$									17.89	
N	5006				4977					

[a] R_W^2 表示群内回归 R^2,R_B^2 表示群间回归 R^2,R^2 表示整个回归 R^2。

对于所有固定效应都相等的零假设，其 F 检验拒绝零假设。固定效应结果基本上具有相同的模型，但是注意到，t 比率甚至更小一些。现在，收入弹性的点估计是 0.60，与之相比，OLS 结果中的点估计则是 0.67。不过，整体而言，对不同变量作用方面的推断并无显著改变。

关于截距上随机变异为 0 的零假设，其 $\chi^2(1)$ 得分检验建立在式(24.43)的基础上，这表明 RE 模型在约束回归方面得到了改进。不过，估计模型也没有导致评价不同变更作用时出现显著变化。正如人们所预期的，在 FGLS 栏目与 RE(GLS)栏目下阐述的结果极为相似。这种较小差异基本上归因于在变换时运用了不同的值。FGLS 估计建立在 $\hat{\rho}=0.12$ 的基础上，该 $\hat{\rho}$ 值是通过对利用最小二乘残差的 100 个估计值再对所获得的 ρ 的 100 个估计值求平均值而得出的。

FE 与 RE 结果上的绝对差异相对很小。非正式比较没有显示，FE 与 RE 公式会产生本质上不同的结果；不过，豪斯曼检验表明，两种估计集合之间存在统计上的显著差异。

总之，这些结果表明，更应该对群间相关性做出某种调整，以及到底怎么做才能拥有对结果相对小的影响。

其次，我们考察利用泊松模型计数变量的结果，这里的计数是指个体到药店的次数。这是一个有意思的变量，因为越南医疗支出的高比例通过购买与使用在药店直接购买的非处方药而采取自述医疗方式。假定这类保健形式比在专业人员指导下所获得的质量更低。在越南，合适个体通常为高收入政府人员与私人部分雇主，他们有能力购买健康保险，享受在政府医院就医，还会获得指定医疗。从表 24.3 发现，样本个体的 16%拥有此类健康保险。

表 24.5 显示 PHARVIS 的观测频数分布，个体中大约 26%在调查期间有不止一次的访问，并且个体中大约 95%有总数为 3 次或更少次数的访问。

表 24.5 越南保健：去药店次数频率

次数	0	1	2	3	4	5	6	7	8	9	10+
PHARVIS	20639	3827	1716	776	359	174	64	43	16	4	115
PHARVIS (分数)	0.744	0.137	0.062	0.028	0.013	0.006	0.002	0.001	0.000	0.000	0.004

表 24.6 阐述了泊松回归的几种变形结果，类似于线性回归的表 24.4 中的那些结果。第一栏给出了泊松 MLE 估计，而普通未调整 t 比率列在第二栏。接下来的栏表示建立在异方差性——一致方差估计基础上的稳健 t 比率。这些都更小一些，在某些情况下，与未调整的那些相比，超过 2 倍，第四栏给出调整整群的建立在利用式(24.45)计算出的方差基础上的 t 比率。第四栏远远小于前面两栏的事实证实了，确实存在群间相关性。平均整群容量大于 140 个观测值；因此，甚至群间相关性程度不大，可能使 t 比率大大增大，而且结果证实了这一点。

其次，我们考察利用 FE 与 RE 模型对群间相关性进行建模。FE 模型可利用条件 MLE 加以估计。省略那些群间变异并不充分的群。得到的估计系数远远不同于通过泊松 MLE 估计所获得的结论。注意到，ln(HHEXP)的系数从显著正的

转变成显著负的。这意味着，最初回归建议去药店次数是一种正态商品，但 FE 估计建议，它是一种劣等商品：也就是说，当收入提高时，个体避免这种自我医疗形式。有理由将此作为固定效应，捕获那些与观测结果相关的省略变量的影响。省略变量可能是对社区居民来说可以利用的另一种医疗服务的数量与质量。这些都可能出现变化，极其依赖于社区的地理区位与经济状况。

表 24.6 越南保健：去药店次数的 RE 模型与 FE 模型

	泊松模型		异方差稳健的	群稳健的	固定效应泊松		随机效应泊松	
变量	系数	|t|	|t|	|t|	系数	|t|	系数	|t|
CONS	−1.637	35.78	18.81	12.25	—	—	1.318	19.41
LNHHEXP	0.78	5.68	3.08	1.90	−0.114	6.01	−0.095	4.95
INSURANCE	−0.245	9.57	5.68	4.29	−0.163	6.17	−0.178	6.44
SEX	0.084	4.96	2.76	2.73	0.098	5.75	0.099	5.71
AGE	0.024	2.38	1.27	1.06	0.03	0.32	0.005	0.55
MARRIED	0.124	5.92	2.96	2.78	0.164	7.59	0.158	7.38
ILLDAYS	0.042	40.00	14.91	12.91	0.046	40.14	0.046	40.18
ACTADYS	0.008	1.71	0.43	0.45	0.025	4.53	0.024	4.35
INJURY	0.171	2.30	0.84	0.85	0.144	1.80	0.143	1.80
ILLNESS	0.562	87.15	24.60	21.81	0.584	73.45	0.585	74.16
EDUC	−0.052	11.10	6.47	3.92	−0.24	4.18	−0.026	4.61
−ln L	25 281				22 446		23 419	
N	27 765				27 671		27 765	

表 24.6 的最后两列给出了建立在随机效应公式基础上的结果。这里假定，泊松分布的截距随不同群而随机变化，每一个群都从共同单变量分布尤其是具有单位均值的伽玛分布中“采样”其截距。这种方式引人注目，其原因是它不要求任何条件。豪斯曼等人（Hausman et al.，1984）研究了一种截距服从伽玛分布的 RE 泊松面板模型，该模型具有解释似然函数，这可以适合于整群数据的情形。对 RE 模型估计得到的结果在性质上类似于对 FE 模型估计所获得的。不过，重要的收入变量估计系数是由在简单泊松假设下所获得的估计经过一番变动得到的。

该例子表明，群间相关性可能产生影响，它不仅对效率有影响，而且对估计值自身也有影响。

24.8 复杂调查

前面几节对分层、加权以及仅有集群内容进行了讨论。本节关注运用分层多阶段整群抽样设计的复杂调查。此类调查的目的是阐述当总体参数可能随不同层而变化时对总体的概括。于是，就要使用加权估计量，并将其看成普查系数的一种估计。一旦控制了可能比 24.5 节更为复杂的集群，目标是一致地估计出加权估计量的方差。

24.8.1 复杂调查的方差估计

我们考察下述设置结构。样本中的第 i 个观测值是位于 s 层第 c 个群内的 j 住户。例如，因变量用 y_{scj} 表示，尽管更正式地讲，可将观测值(s,c,j)重新表述成观测值(s,c_s,j_{c_s})。数据是$(y_{scj},\mathbf{x}_{scj},w_{scj})$，其中，$w_{scj}$ 是一个反比例于选取样本观测值概率的样本权数。下标利用了非汇总水平加以排序，与 24.5 节中的记号相反。

二级抽样[1](two-stage)或多级抽样[2](multistage sampling)用于层内，所选取的住户作为至少两个序贯采样的结果。首先，该层内的所有 PSU 的子集是随机抽取的。其次，对选取 PSU 中的所有住户抽取子集，其中允许出现整群抽样。进一步地，还可能从 SSU 内采样。

线性统计量方差

起点是考察线性统计量的方差估计，这里的线性统计量是关于层、PSU 以及住户的求和：

$$\hat{u}=\sum_{s=1}^{S}\sum_{c=1}^{C_s}\sum_{j=1}^{N_{cs}}u_{scj}=\sum_{s=1}^{S}\sum_{c=1}^{C_s}u_{sc}$$

其中，u_{sc} 表示 PSU 之内的全体，因而：

$$u_{sc}=\sum_{j=1}^{N_{cs}}u_{scj}$$

下面将给出 u_{scj} 的例子，诸如加权均值与加权回归。u 的方差是：

$$\mathrm{V}[u]=\sum_{s=1}^{S}\sum_{c=1}^{C_s}\mathrm{V}[u_{sc}]=\sum_{s=1}^{S}C_s\sigma_s^2$$

若我们假定 u_{sc} 关于层是独立的且关于 PSU 是 iid 的，具有共同方差 σ_s^2。已知 u_{sc} 关于 c 为 iid 的，可以利用 σ_s^2 的通常无偏方差估计值，所以 $\hat{\sigma}_s^2=(C_s-1)^{-1}\sum_c(u_{sc}-\bar{u}_s)^2$。由此可得：

$$\mathrm{V}[\hat{u}]=\sum_{s=1}^{S}\frac{C_s}{C_s-1}\sum_{c=1}^{C_s}(u_{sc}-\bar{u}_s)^2 \tag{24.55}$$

其中，$\bar{u}_s=C_s^{-1}\sum_c u_{sc}$ 表示 PSU 全体的层平均值。

这个估计量考虑到了内部的集群，因为：

$$\begin{aligned}\sum_{c=1}^{C_s}(u_{sc}-\bar{u}_s)^2&=\sum_{c=1}^{C_s}\Big(\sum_{j=1}^{N_{cs}}u_{scj}-\bar{u}_s\Big)^2\\&=\sum_{c=1}^{C_s}\sum_{j=1}^{N_{cs}}(u_{scj}-\bar{u}_s)^2+\sum_{c=1}^{C_s}\sum_{j=1}^{N_{cs}}\sum_{k\neq j}^{N_{cs}}(u_{scj}-\bar{u}_s)(u_{sck}-\bar{u}_s)\end{aligned}$$

第一个和式是在 SRS 条件下对方差的贡献。第二个和式在整群抽样下将是正的，从而引起较大方差。若不对抽样特性做出假设，就不会产生层内聚集形式。例如，式(24.55)给出正确的标准误差，即使存在三级抽样，而且针对 SSU 有进一步二次

[1] 又称为二阶抽样。—— 译者注

[2] 又称为多阶抽样。—— 译者注

抽样。

估计量式(24.55)确实要求,至少两个 PSU 是从每一个层抽取的。当仅有一个 PSU 被采样时,则一种可能性是要合并一些层,包括单个 PSU 进入另一个层中,将此层看成类似于一个合理的先验。一种可行情况是,倘若 $C_s \geq 2$,即每层至少存在两个 PSU。由于不同层出现各不相同的均值,当引入向上偏倚时,这将导致对 $V[u]$ 的高估。[1]

在实际应用中,PSU 都是以不放回方式抽样的,因此,u_{sc} 中存在某种相依性。从而,类似于 24.2.3 节情况,式(24.55)高估了 $V[u]$。为此,人们提出了更复杂的公式。

加权均值的方差

总体均值可通过 y_{scj} 的样本加权总数(比如说 $\hat{y}$)与样本权数之和(比如说 $\hat{w}$)的比值来加以估计。于是:

$$\bar{y}_W = \hat{y}/\hat{w} = \sum_{s=1}^{S}\sum_{c=1}^{C_s}\sum_{j=1}^{N_{cs}} w_{scj} y_{scj} \Big/ \sum_{s=1}^{S}\sum_{c=1}^{C_s}\sum_{j=1}^{N_{cs}} w_{scj}$$

当将样本权数看成已知的,则有更简单的形式:

$$\bar{y}_W = \sum_{s=1}^{S}\sum_{c=1}^{C_s}\sum_{j=1}^{N_{cs}} w_{scj}^{*} y_{scj}$$

其中,$w_{scj}^{*} = w_{scj}/\hat{w}$,利用满足 $u_{scj} = w_{scj}^{*} y_{scj}$ 的式(24.55),就可应用 $V[\bar{y}_W]$。

若将样本权数处理成未知的,则运用德尔塔方法或线性化方法得出:$V[\hat{y}/\hat{w}]$ 作为 $V[\hat{y}]$、$V[\hat{w}]$、$\mathrm{Cov}[\hat{y},\hat{w}]$ 的一种函数。前面这两个量能利用满足 $u_{scj} = w_{scj} y_{scj}$ 且 $u_{scj} = w_{scj}$ 的式(24.55)估计出来。第三个量可通过用 $(u_{sc} - \bar{u}_s)(v_{sc} - \bar{v}_s)$ 代替式(24.55)中的 $(u_{sc} - \bar{u}_s)^2$ 而得到估计,其中,$u_{scj} = w_{scj} y_{scj}$ 且 $v_{scj} = w_{scj}$。这是比值估计量的一个例子。

对于非线性统计量,诸如这些比值估计,文献已经提出了基于刀切法(**jack-knife**)或平衡重复复制。由于非线性原因,方差估计不再是无偏的,但可以证明,当层数 $S \to \infty$ 时,它是一致的[参见克鲁斯和拉奥(Krewski and Rao, 1981)]。沃尔特(Wolter, 1985)对 S 固定且 $\sum_{c=1}^{C_s} N_{cs} \to \infty$ 时的某些结果进行了总结。人们还能实施自助法,尽管运用时需要小心谨慎。参见拉奥和吴建福(Rao and Wu, 1988),以及绍和图(Shao and Tu, 1995)。

加权最小二乘估计量方差

由 24.3 节知,普查回归系数的加权回归估计值 $\hat{\boldsymbol{\beta}}_W$ 是

$$\sum_{s=1}^{S}\sum_{c=1}^{C_s}\sum_{j=1}^{N_{cs}} w_{scj} \mathbf{x}_{scj} (y_{scj} - \mathbf{x}_{scj}' \hat{\boldsymbol{\beta}}_W) = \mathbf{0}$$

---

[1] 对于 CPS,这里的方法并不能直接应用,因为许多层仅有一个 PSU,而就其他层而言,只有一个 PSU 被收集。不过,各种伪层可以建立起来,并运用一些从伪层中重复抽取 PSU 的复制方法。参见美国人口普查局(2002)。——译者注

的解。经过一些通常代数运算，得到：

$$\hat{\boldsymbol{\beta}}_{\mathrm{W}}-\boldsymbol{\beta}=\left(\sum_{s=1}^{S}\sum_{c=1}^{C_s}\sum_{j=1}^{N_{cs}}w_{scj}\mathbf{x}_{scj}\mathbf{x}_{scj}'\right)^{-1}\times\sum_{s=1}^{S}\sum_{c=1}^{C_s}\sum_{j=1}^{N_{cs}}w_{scj}(y_{scj}-\mathbf{x}_{scj}'\hat{\boldsymbol{\beta}}_{\mathrm{W}})$$

这得到了三明治形式方差 $\mathrm{V}[\hat{\beta}]=\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$，其中，$\mathbf{B}$ 表示第二个三重求和的方差，这能利用满足 $u_{scj}=w_{scj}\mathbf{x}_{scj}(y_{scj}-\mathbf{x}_{scj}'\hat{\boldsymbol{\beta}}_{\mathrm{W}})$ 的式(24.55)估计出来。

加权 m 估计量的方差

一种相当一般的框架考察加权 m 估计量 $\hat{\boldsymbol{\theta}}_{\mathrm{W}}$，它是

$$\sum_{s=1}^{S}\sum_{c=1}^{C_s}\sum_{j=1}^{N_{cs}}w_{scj}\mathbf{h}(y_{scj},\mathbf{x}_{scj},\hat{\boldsymbol{\theta}}_{\mathrm{W}})=\mathbf{0}$$

的解。例子包括，线性回归 $\mathbf{h}_{scj}=\mathbf{x}_{scj}(y_{scj}-\mathbf{x}_{scj}'\boldsymbol{\beta})$，以及拟极大似然 $\mathbf{h}_{scj}=\partial\ln f(y_{scj}\mid\mathbf{x}_{scj},\boldsymbol{\theta})/\partial\boldsymbol{\theta}$。

一旦假定 $\boldsymbol{\theta}$ 有一致估计，这要求 $\mathrm{E}[\mathbf{h}(y_{scj},x_{scj},\boldsymbol{\theta})]=\mathbf{0}$，我们能使用估计方程的通常一阶泰勒级数表达式，得到：

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{\mathrm{W}}-\boldsymbol{\theta})\xrightarrow{d}\mathcal{N}[\mathbf{0},\mathbf{A}^{-1}\mathbf{B}\mathbf{A}'^{-1}]$$

其中：

$$\mathbf{A}=\operatorname{plim}N^{-1}\sum_{s=1}^{S}\sum_{c=1}^{C_s}\sum_{j=1}^{N_{cs}}w_{scj}\frac{\partial\mathbf{h}(y_{scj},\mathbf{x}_{scj},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}$$

与：

$$\mathbf{B}=\operatorname{plim}N^{-1}\sum_{s=1}^{S}\sum_{c=1}^{C_s}\sum_{j=1}^{N_{cs}}\sum_{k=1}^{N_{cs}}w_{scj}w_{sck}\mathbf{h}(y_{scj},\mathbf{x}_{scj},\boldsymbol{\theta})\frac{\partial\mathbf{h}(y_{sck},\mathbf{x}_{sck},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}$$

其中，假定 $\mathbf{B}$ 的表达式与 $\mathbf{h}_{scj}$ 在层与整群上均是独立的，但允许在层内出现相依性。对 $\mathbf{A}$ 的估计可直接获得。对于 $\mathbf{B}$ 来说，运用满足 $u_{scj}=w_{scj}\mathbf{h}_{scj}$ 的式(24.55)，得到：

$$\mathbf{B}=\sum_{s=1}^{S}\frac{C_s}{C_s-1}\sum_{c=1}^{C_s}[\bar{z}_{sc}-\bar{z}_s]^2$$

其中，$\bar{z}_{sc}=\sum_{j=1}^{N_{cs}}w_{scj}\mathbf{h}(y_{scj},\mathbf{x}_{scj},\boldsymbol{\theta})$，$\bar{z}_s=C_s^{-1}\sum_{c=1}^{C_s}\bar{z}_s$。

内生分层

左方(Sakata，1998)将这些结果推广到内生抽样。他采用了普查参数方法，并给出了在假定层数目 $S\rightarrow\infty$ 时的渐近理论。其结果与上一节讨论的那些结果一样。

24.9 应用研究

在微观经济计量学研究中，采用结构方法最为普遍。倘若不存在内生分层，则可使用加权估计量。如果存在聚集，主要关注内容是获得正确的标准误差。当整群效应是随机的时候，估计中忽略聚集的有效性一般损失很少。一些软件包可能拥有整群稳健标准误差选项，不要与异方差性稳健选项相混淆，假如整群效应是随

机的，并且存在众多群，则运用整群稳健标准误差是适宜的。倘若在 CSFE 情况下，存在并不太多的群，则可利用 OLS 实施 CSRE 与 CSFE 模型。否则，使用面板数据模块，如果该模块支持非平衡面板。对于面板数据来说，非经济计量学领域的大多数研究者对采用随机效应方法满意，可是为了一致估计，可能必须采用固定效应方法。

若采用描述性方法，并且参数随不同层而变化，则必须进行加权。在最小二乘法内部运用加权选项，但它必须与整群稳健标准误差选项相结合。一些软件包具有调整估计模块，该模块利用 24.6 节方法获得整群标准误差。软件包 SUDAAN 能执行本章中的线性回归模型与重要的非线性回归模型。

24.10 文献注释

24.2－24.3 抽样调查文献极为丰富。抽样调查方面的经典参考书包括基什(Kish，1965)、科克伦(Cochran，1977，1953 年第 1 版)。斯金纳(Skinner，1989)给出了一个有用的综述，格罗夫斯(Groves，1989)提供了一种相对非技术性研究，阐述社会科学进行调查的众多方法，以及产生的许多有用的实际问题。为了完整起见，我们包含了某些这类抽样调查文献，尽管经济计量学研究很少运用 24.8 节的方法。除了著名的帕德尼(Pudney，1989)、迪顿(Deaton，1997)书中的一些章节以及乌拉和布罗伊宁(Ullah and Breuning，1998)的书之外，还有少数的经济计量学文献。

24.4 理论经济计量学文献的主要焦点是控制内生分层。这方面的文献具有挑战性，我们只提供一个概述。详细内容参见雨宫(Amemiya，1985)，他提供了许多文献，其中包括曼斯基和莱尔曼(Manski and Lerman，1977)的离散选择模型，以及豪斯曼和怀斯(Hausman and Wise，1979)的样本选择模型。尽管简单加权估计量无效，但一般地讲，它是适宜的。英伯斯和兰开斯特(Imbens and Lancaster，1996)阐述了在已知条件密度下，实施完全有效估计量的实用方法。

24.5 对于微观经济计量学应用来说，控制聚集是极为重要的。克勒克(Kloek，1981)与莫尔顿(Moulton，1986，1990)的研究工作是促使经济计量学家改变此问题的一个关键。戴维斯(Davis，2002)给出了多重方式误差成分方法的一种一般性研究。格劳巴德和科恩(Graubard and Korn，1994)曾提出整群数据线性回归分析的一个有益讨论。他们既关注固定效应模型，又考虑随机效应模型，强调了使随机效应模型成为有效的假设必须被满足。彭德格斯特等人(Pendergast et al.，1996)给出分析整群二值数据方法的广泛综述。由于式(23.34)右边的中间项涉及对整群数目求均值，这种估计的准确性依赖于整群数目。当群数目很小时，利用整群稳健方差矩阵的结果仍是一个值得探索的专题[唐纳德和朗(Donald and Lang，2001)；安格里斯特和拉维(Angrist and Lavy，2002)]。伍德里奇给出了一个综述(Wooldridge，2003)。

24.6 社会科学中广泛运用分层线性模型。布雷克和劳登布什(Bryk and Raudenbush，2002)既从似然观点又从贝叶斯观点，给出一种综合涵盖二值结果、

有序结果、计数结果以及多项式结果的论述。

24.7 世界银行对发展中经济实施了各种生活水平调查，迪顿(Deaton，1997)运用来自世界银行的整群样本数据讨论了一系列建模型问题。

24.8 许多标准统计软件包比如STATA与SUDAAN，针对横截面数据与面板数据，提供了线性和非线性模型中的固定效应公式以及随机效应公式。

习 题

24-1 (a) 验证由式(24.25)给出的 Σ_c 表达式。

(b) 证明CSRE模型的估计量 $\hat{\sigma}^2$ 与 $\hat{\rho}$ 具有一致性。

(c) 考察平衡整群CSRE模型的标准误差偏倚。证明在此情况下，$\mathrm{E}[\sum_c\sum_j\hat{u}_{cj}^2]=\sigma^2[N-K(1+\rho(m-1))]$。

24-2 [改编自格林沃尔德(Greenwald，1983)。]考察线性回归模型 $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{u}$，其中，$\mathrm{E}[\mathbf{u}]=\mathbf{0}$，并且 $\mathrm{E}[\mathbf{uu}']=\sigma^2\boldsymbol{\Omega}^*=\boldsymbol{\Omega}$。利用OLS估计量 $\hat{\boldsymbol{\beta}}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$的标准结果(参见4.4节)，我们能获得 $\mathrm{V}[\hat{\boldsymbol{\beta}}]$的正确表达式，因为$\mathbf{V}_2=(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})^{-1}$，而 $\mathbf{V}_1=\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$，当 $\boldsymbol{\Omega}\neq\mathbf{I}$ 时，$\hat{\sigma}^2=\hat{\mathbf{u}}'\hat{\mathbf{u}}/(N-K)$是无效的。

(a) 证明 $\mathbf{V}_1$ 的偏倚由 $\mathbf{B}=\mathbf{B}_1+\mathbf{B}_2$ 给出，其中，$\mathbf{B}_2=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\boldsymbol{\Omega}-\sigma^2\mathbf{I})\mathbf{X}\times(\mathbf{X}'\mathbf{X})^{-1}$，而 $\mathbf{B}_1=(N-K)^{-1}\mathrm{tr}\{\mathbf{B}_2(\mathbf{X}'\mathbf{X})\}(\mathbf{X}'\mathbf{X})^{-1}$。(格林沃尔德将 $\mathbf{B}_2$ 称为“直接偏倚”。)

(b) 对于特殊情况 $\mathbf{X}'\mathbf{X}=\mathbf{I}_K$，计算两项。证明当 $N\rightarrow\infty$时，$\mathbf{B}\rightarrow\mathbf{B}_2$。

24-3 考察OLS整群稳健方差估计量公式(24.33)。假定存在两个水平聚集。具体地讲，在本章实证例子背景下，聚集能在家庭与社区水平上出现，如果来自相同社区家庭的多位成员都在调查之中。当数据有两种水平聚集时，该公式将怎样进行修改?

24-4 对于这个习题，运用VLSMS数据的50%样本。当实验者至少有一次去药店(PHARVIS)时，定义 $y=1$，否则定义 $y=0$。本题假定可以运用处理集群的程序。

(a) 使用的解释变量与24.7节泊松模型中的那些一样，既用到方差标准估计量，又用到方差稳健三明治估计量，通过极大似然法对二值logit模型加以估计。

(b) 利用整群稳健标准误差选项，重新对(a)部分设定进行估计。解释(a)部分与(b)部分的稳健标准误差之间的差异。

(c) 运用“社区”作为整群标志符。用整群固定效应与整群随机效应设定，重新估计logit模型。对LNHHEXP与INSURANCE的系数估计及标准误差进行比较。这两个变量的显著性结论会受到数据聚集影响吗?

25 处理评估

25.1 引　论

处理评估专题涉及干预对关注结果的影响进行测算，对干预和结果类型可以宽泛地加以定义，以便应用于众多不同背景的内容上。处理评估方法及其某些术语均源自医学，其中，干预经常意指采用的处理体系。因此，人们可能对测量与某一基准——例如没有处理或不同处理——有关的处理响应感兴趣。在经济应用中，处理与干预常常意指同一个含义。

在经济背景下，处理的例子包括劳动力培训项目注册、成为贸易联盟的成员、接收来自社会项目的调动、接收来自社会项目制度方面的变化、关于金融交易规则与制度方面的变化、经济激励变动等；参见莫菲特（Moffitt, 1992），弗里德伦德、格林伯格和罗宾斯（Friedlander, Greenberg, and Robbins, 1997），以及赫克曼、拉隆德和史密斯（Heckman, Lalonde, and Smith, 1999）的文献。如果所使用的处理能够随着强度或类型不同而变化，当对它们进行汇总研究时，我们就用多重处理（**multiple treatments**）这一术语。与单一类型处理有关，这并没有引致复杂性，但现在为了对此进行研究，对基准的选择更为灵活。

结果术语表示经济地位或者个体的经济状况环境变化。一种重要情况是，当关注结果是连续变量的情形，比如说 y，而处理变量是离散的且处于变化/不变化，比如说 D，如果处理得到应用，那么 D 取值 1，否则 D 取值 0。干预的一个例子是，劳动力市场培训，这种培训能影响到工人培训后的工资。然而，通常结果要么是连续的或离散的，要么表现出受限变化。可是，详细分析将会随情况不同而变化，但某些重要思想在所有情况中都是有意义的。为了简单起见，将连续结果与二值处理作为我们研究的主要情况。稍后，将这种分析扩展到其他有关的特别情形。

处理评估的政策意义是直接的，因为“成功”处理与人们期望的社会项目相联系，或者已有项目的改进达到社会政策等目标。赫克曼与史密斯（Heckman and Smith, 1998）曾经讨论过，几种广泛使用的测量处理影响以及与传统成本效益分析之间的关系。

处理评估的标准问题包括对处理与结果之间的因果联系进行推断。在标准单个处理例子中，我们可观测到$(y_i, \mathbf{x}_i, D_i)$，$i=1,\cdots,N$，而一旦 $\mathbf{x}$ 保持固定不变，关

于 y 的假设 D 变动的影响是人们关注的内容。这种推断是潜在结果模型的重要特征，已在第 2 章引进，在此情况下，关注结果变量就可在已处理状态与未处理状态之间加以比较。不过，不是所有个体在这两种状态中都可以同时被观测到。因此，该情况接近于缺失数据，但它能借助于反事实(**counterfactuals**)所完成的因果推断方法加以研究。倘若一个人接受处理，则将探索平均处理个体的结果会出现怎样变动。也就是说，关注诸如数量 $\Delta y/\Delta D$。人们的关注内容本质上是由这种干预引起的。此处，因果是在其他条件相同(**ceteris paribus**)的意义下，对所有其他变量保持常值。

本章和前面几章之间的区别是什么呢？对此，我们还会考察各种模型的识别与估计吗？它们存在许多相似点，但其差异源自强调内容的变动。主要差异来源于对处理有效性的测量族。这些测量都是参数与数据的函数，同时它们能够比较有关政策的反事实。一个重要而有意思的结果是，已知数据与估计量，并不能建立所有的测量。在估计模型时，对所使用的估计量与数据类型的选择均受限于能够成为可识别的反事实，从而能一致估计出影响测量。

在处理评估文献中，另一个强调内容是，保证利用最小函数形式与排除约束的识别优点(例如，半参数识别)。这种强调是由产生政策意义但其有效性并不依赖于强假设愿望而引发的。半参数识别的可行性，在含有关于因变量连续支集的线性模型中，建立处理效果估计，与在含有受限因变量的非线性模型中建立处理效果估计相比，相对更容易一些。

25.2 节讨论识别性假设。25.3 节阐述处理效果的测量，这通常是识别与估计的目标。25.4 节分析匹配估计量与倾向得分估计量。25.5 节涵盖处理效果的差异中差分估计量，这是在拟试验数据设置背景下研究事件所普遍采用的。

一旦继续拥有拟试验设置背景，在 25.6 节，我们讨论回归非连续性设计，然后在 25.7 节借助于工具变量估计量进行研究。迄今为止，大多数讨论内容都与线性模型有关。而 25.8 节提供运用本章介绍的方法进行详细阐明的一个例子。

25.2　背景设置与假设

对处理效果进行估计的方法依赖于促使因果效果可识别的一些假设，例如，线性 SEM 依赖于允许因果效果的假设(参见第 2 章)。在本节，我们详述允许使用重要匹配估计量与倾向得分估计量的假设，这些估计量稍后在 25.4 节加以阐述。

首先，研究实施估计时对因果参数进行估计的框架。

25.2.1　*处理效果框架*

让我们以社会实验中对处理指派的随机化设置背景开始，如同 3.3 节所阐述的。设存在关注处理的目标总体，并设 N 表示随机选取个体数目，这些个体是自愿参与处理的。设 N_T 表示随机选取的已处理个体数，设 $N_C=N-N_T$ 表示未处理个体数，它们作为潜在对照组[1](**control group**)。

---

[1] 又称为控制组。——译者注

随机指派蕴含着，指派会忽略处理对结果的可能影响。例如，在处理组中，不存在由于个体预期利益大而被列在处理范围之内，同时因个体预期利益小则不予考虑的这种情况。设$(y_i, \mathbf{x}_i, D_i;\ i=1,\cdots,N)$表示关于纯量值结果变量(**outcome variable**)观测值 y 的向量、可观测变量 $\mathbf{x}$ 的向量以及处理变量 D 的二值指示变量。为了简单起见，假定被指派处理的任何人都参与，而没有被指派处理的任何人都不参与。已处理个体的结果变量记为 y_1，而未处理个体的结果变量记为 y_0。在实施试验并且搜集到数据之后，我们希望获得处理影响的测算。对处理效果进行测量的一种最普通方法是，建立对已处理(**treated**)平均结果组与未处理(**nontreated**)组平均结果比较的测算。

与之相伴的一个重要差异是，相同数据设置背景能用于观测数据。该差异在于不存在对处理的随机指派机制。或许因为个体已被选取处理，或者因为某种其他原因。

开始时就需要声明，绝大部分处理评估研究具有部分平衡特征。具体地讲，人们假定不存在一般平衡效果。由此，我们意指处理效果小且不会影响到某些被看成是外生的变量状况。如果人们考察会影响整个部门的处理项目，而该部分是国民经济的重要部分，那么这个假设将不成立。例如，设立全体健康保险会对整个健康服务部分产生影响，这很难应用本章讨论的方法。

在建立处理效果估计时，存在许多潜在陷阱。建立这类测算的假设变化而引起的各种解释之间存在着微妙差异。因此，我们通过审视这些假设来开始。

25.2.2 *条件独立性假设*

对两个组结果之间进行有意义的比较，需要某些假设。我们首先列出并解释这些假设，稍后在讨论某个处理效应的可识别性时，使用它们。

一个重要的假设是条件独立性假设(**conditional independence assumption**)，其内容表述如下，以 $\mathbf{x}$ 为条件的结果与处理是独立的，可写成：

$$y_0, y_1 \perp D \mid \mathbf{x} \tag{25.1}$$

该假设的行为含义是指一旦控制由 $\mathbf{x}$ 不同而引起的结果差异之后，处理项目的参与不依赖于结果。正确运用随机指派，将会证实这个假设。实际上，在完全随机指派下，人们甚至可做出一个较强假设：

$$y_0, y_1 \perp D \tag{25.2}$$

因为随机化是针对$(y, \mathbf{x})$空间进行的。更广泛使用假设(25.1)，假如该假设有效，它对某些影响参数的识别是有用的，因为它表述的内容是，一旦我们控制住某些与 D 有关的回归元 $\mathbf{x}$ 的效应，则处理与结果均是独立的。

条件独立性假设具有很宽泛的意义，并蕴含着下述内容：

$$F(y_j \mid \mathbf{x}, D=1) = F(y_j \mid \mathbf{x}, D=0) = F(y_j \mid \mathbf{x}),\quad j=0,1 \tag{25.3}$$
$$F(u_j \mid \mathbf{x}, D=1) = F(u_j \mid \mathbf{x}, D=0) = F(u_j \mid \mathbf{x}),\quad j=0,1$$

其中，u 表示回归模型误差，它意味着参与决策没有影响到潜在结果的分布(**distri-**

bution of potential outcomes)。

为了理解这个假设的影响，设 $\mathrm{E}[y|\mathbf{x},D]$ 是线性的，也就是说，参与结果方程是：

$$y=\mathbf{x}'\boldsymbol{\beta}+\alpha D+u \tag{25.4}$$

其中，$\mathrm{E}[u|D]=\mathrm{E}[y-\mathbf{x}'\boldsymbol{\beta}-\alpha D|D]=0$。因此，可将 D 处理成外生变量，而且不存在联立性偏倚或选择性偏倚。在标准的以 $\mathbf{x}$ 为条件下，对回归参数进行一致估计是可行的。

比式(25.1)稍弱的一个假设是：

$$y_0\perp D|\mathbf{x} \tag{25.5}$$

这蕴含着参与同 y_0 是独立的。该假设用于建立对已处理总体平均处理效应(**treatment effect on the treated, ATET**)的可识别性，正如稍后将看到的那样。

文献中，假设式(25.5)还有其他称谓。英伯斯(Imbens, 2005)称它为非混乱性假设(**unconfoundedness assumption**)，而鲁宾称它为可忽略性假设(**ignorability assumption**)[鲁宾(Rubin,1978)，伍德里奇(Wooldridge, 2001)]。假如有效，该假设蕴含着一旦被包括在回归中，就没有省略变量偏倚(**omitted variable bias**)，因此将不会出现混淆局面。这个假设与忽略结果的处理指派有关；因而，称它为可忽略性假设是适宜的。

如果将处理变量看成是外生的，就必须含有该假设，为使估计简单，这样做必不可少。如果有效，就不需要样本选择模型或用于处理内生处理变量的 IV 方法，但可运用 25.4 节的方法。

25.2.3 *匹配假设*

第二个假设称为交叉或匹配假设(**overleap** or **matching assumption**)，它是识别影响某种总体测量所必需的。对它表述如下：

$$0<\Pr[D=1|\mathbf{x}]<1 \tag{25.6}$$

该假设确保了，对于 $\mathbf{x}$ 的每一个值，既存在已处理情况又存在未处理情况。在此意义下，在已处理子样本与未处理子样本之间存在着交叉。对于每一个已处理个体来说，存在另一个具有类似 $\mathbf{x}$ 的匹配未处理个体。如果假设失效，那么能潜在地拥有已处理的 $\mathbf{x}$ 向量个体，以及未处理的不同 $\mathbf{x}$ 的那些个体。对于识别已处理组的处理参数来说，就不要求这一假设。对于识别被随机选取的个体的处理效应来说，都需要每一个参加者有一个类似的非参加者。于是，条件 $\Pr[D=1|\mathbf{x}]<1$ 就足够了。

25.2.4 *条件均值假设*

第三个假设是条件均值独立性假设(**conditional mean independence assumption**)：

$$\mathrm{E}[y_0|D=1,\mathbf{x}]=\mathrm{E}[y_0|D=0,\mathbf{x}]=\mathrm{E}[y_0|\mathbf{x}] \tag{25.7}$$

这意味着 y_0 不能决定参与。

25.2.5 倾向得分

当处理参与不是由随机指派的，而是随机地依赖于可观测变量 **x** 向量，如同观测数据一样，或将处理作为某个由一些可观测特征(例如年龄、性别或社会经济地位)所定义的总体时，倾向得分(**propensity scores**)概念是有用的。这是一个 **x** 给定时关于处理参与的条件概率测量，并用 $p(\mathbf{x})$ 表示：

$$p(\mathbf{x})=\Pr[D=1|\mathbf{X}=\mathbf{x}] \tag{25.8}$$

倾向得分测量可在给定数据$(D_i,\mathbf{x}_i)$时，利用第 14 章研究的参数或半参数方法(例如，通过做一个 logit 回归)计算出来。

在处理评估中，起着重要作用的假设是平衡条件(**balancing condition**)，它可表述成：

$$D\perp\mathbf{x}|p(\mathbf{x}) \tag{25.9}$$

这能以另一种可选择的方式表述如下：对于具有相同倾向得分的个体来说，指派处理是随机的，从而它们的 **x** 向量应该看起来是一样的。平衡条件是一个可检验的假设。

给定 $p(\mathbf{x})$ 时条件独立性的一种有用结果是由罗森鲍姆和鲁宾(Rosenbaum and Rubin, 1983)给出的，它可表述成：

$$y_0,y_1\perp D|\mathbf{x}\Rightarrow y_0,y_1\perp D|p(\mathbf{x}) \tag{25.10}$$

这蕴含着，给定 **x** 时的条件独立性假设意味着给定 $p(\mathbf{x})$ 时存在条件独立性，也就是说，给定 $p(\mathbf{x})$ 时，y_0、y_1 以及 D 都是独立的。

为了获得这一结果，注意到：

$$\begin{aligned}\Pr[D=1|y_0,y_1,p(\mathbf{x})]&=\mathrm{E}[D|y_0,y_1,p(\mathbf{x})]\\&=\mathrm{E}[\mathrm{E}[D|y_0,y_1,p(\mathbf{x}),\mathbf{x}]|y_0,y_1,p(\mathbf{x})]\\&=\mathrm{E}[\mathrm{E}[D|y_0,y_1,\mathbf{x}]|y_0,y_1,p(\mathbf{x})]\\&=\mathrm{E}[\mathrm{E}[D|\mathbf{x}]|y_0,y_1,p(\mathbf{x})]\\&=\mathrm{E}[p(\mathbf{x})|y_0,y_1,p(\mathbf{x})]\\&=p(\mathbf{x})\end{aligned}$$

这里，第二行与第三行均用到了期望迭代定律。第四行等式利用了条件独立性。支持此结果的一种直觉是，$p(\mathbf{x})$是 **x** 的一种特殊函数，在某种意义上，$p(\mathbf{x})$包含的信息比 **x** 中的要少一些。因此，给定 $p(\mathbf{x})$时，条件独立性必然包含给定 **x** 时的相同内容。因为通过以 **x** 为条件，可去掉 **x** 与 D 之间的相关性，同样地通过以倾向得分 $p(\mathbf{x})$为条件，也可去掉 **x** 与 D 之间的相关性。因而，类似于式(25.4)的回归是：

$$y=\mathbf{x}'\boldsymbol{\beta}+\alpha p(\mathbf{x})+u \tag{25.11}$$

$$=\mathbf{x}'\boldsymbol{\beta}+\alpha\hat{p}(\mathbf{x})+(u+\alpha(p(\mathbf{x})-\hat{p}(\mathbf{x}))) \tag{25.12}$$

其中，第二行中未知 $p(\mathbf{x})$可用样本估计量来代替，导致了除抽样误差以外还有回归误差。这种策略的优缺点稍后将加以考虑。表 25.1 对记号及含义给出了一个

总结。

表 25.1　处理效应框架

符号	定　义
y_1	已处理组结果
y_0	未处理组结果
$p(\mathbf{x})$	倾向得分
N_T	样本中处理案例个数

25.3　处理效应与选择偏倚

我们以阐述两个广泛运用的处理效应测量开始，这两个测量中，一个是对所有个体进行平均，而另一个是仅仅对已处理个体进行平均。然后，我们以某种详细方式讨论选择对处理的作用。25.4～25.6 节阐述一些方法，假定选择效果直接依赖于个体的唯一可测量特征，比如年龄。此外，如果选择效果依赖于不可观测成分，就必须使用第 16 章的方法。本节包括对选择问题的重要讨论。

25.3.1　*两个重要参数：ATE 与 ATET*

将 Δ 定义成已处理个体与未处理个体结果之差，它可表述成：

$$\Delta = y_1 - y_0 \tag{25.13}$$

此处，假如愿意，还可以以 **x** 为条件表述。需要强调的是，Δ 是不能直接观测到的，因为没有一个个体能在两个状态下均被观测到。将平均处理效应（**average treatment effect**，简记为 ATE）与已处理的平均处理效应（**average treatment effect on the treated**）的总体值定义成：

$$\text{ATE} = \text{E}[\Delta] \tag{25.14}$$

$$\text{ATET} = \text{E}[\Delta | D = 1] \tag{25.15}$$

其样本类似形式为：

$$\widehat{\text{ATE}} = \frac{1}{N}\sum_{i=1}^{N}[\Delta_i] \tag{25.16}$$

$$\widehat{\text{ATET}} = \frac{1}{N_T}\sum_{i=1}^{N_T}[\Delta_i | D_i = 1] \tag{25.17}$$

其中，$N_T = \sum_{i=1}^{N} D_i$。就这两种情况的每一个而言，若能获得 Δ_i，则可直接进行计算。由于公式含有必须加以估计的不可观测成分，并且估计步骤要求某些假设，所以该方法并不能直接运用。

当处理具有普适应用性，ATE 测量才有意义，所以对随机选取的总体成员来说，考察从处理中获得的假设增益是合情合理的。当我们考察已处理个体从处理中获得的平均增益时，ATET 测量才有意义。参见赫克曼和维特拉西尔（Heckman and Vytlacil，2002）。

为了理解处理评估问题，考察给定特征 $\mathbf{x}$ 时从参与中获得的增益平均。这就是：

$$\begin{aligned} \text{ATE} &= \text{E}[\Delta \mid X=\mathbf{x}] \\ &= \text{E}[y_1 - y_0 \mid X=\mathbf{x}] \\ &= \text{E}[y_1 \mid X=\mathbf{x}] - \text{E}[y_0 \mid X=\mathbf{x}] \\ &= \text{E}[y_1 \mid \mathbf{x}, D=1] - \text{E}[y_0 \mid \mathbf{x}, D=0] \end{aligned} \tag{25.18}$$

其中，最后一个等式使用了条件独立性假设(25.1)。

给定参加者样本，可以对 $\text{E}[y_1 \mid D=1, \mathbf{x}]$进行估计。然而，$\text{E}[y_0 \mid \mathbf{x}, D=0]$是不可观测的，因为它是对那些事实上没有参与的参加者平均结果的测量，同时人们不能同时观测到同一个个体既是参加者又是非参加者。为了执行 ATE 运算，就必须求出第二项估计量。

由定义(25.18)知：

$$\begin{aligned} \text{ATE} &= \text{E}[y_1 \mid \mathbf{x}, D=1] - \text{E}[y_0 \mid \mathbf{x}, D=0] && (25.19) \\ &= \mu_1(\mathbf{x}) - \mu_0(\mathbf{x}) + \text{E}[u_1 \mid \mathbf{x}, D=1] - \text{E}[u_0 \mid \mathbf{x}, D=0] \\ &= \mu_1(\mathbf{x}) - \mu_0(\mathbf{x}) + \text{E}[u_1 \mid \mathbf{x}] - \text{E}[u_0 \mid \mathbf{x}] \\ &= \mu_1(\mathbf{x}) - \mu_0(\mathbf{x}) && (25.20) \end{aligned}$$

其中，等号右边第一行中第一项可利用源自处理参加者的数据得到估计，而第二项却不能直接观测到。第三行用到了条件独立性与条件均值假设，同时对已处理个体采用 $y_1 = \mu_1(\mathbf{x}) + u_1$ 设定，并对未处理个体采用 $y_0 = \mu_0(\mathbf{x}) + u_0$ 设定。最后一行第二项仅仅需要均值独立性，而不是完全条件独立性。

25.3.2 *抽样偏倚与选择偏倚*

评估问题的症结为，$\text{E}[y_0 = \mathbf{x}, D=1]$是不可观测的。对此问题的解决部分依赖于可利用数据的类型。社会实验都使用适宜参加者，而这些参加者被排除在组之外作为反事实的代表。观测研究从相同资源或从其他数据库中生成对照组(**comparison group**)作为已处理组，同时以利用 $\text{E}[y_0 \mid \mathbf{x}, D=0]$的某个函数来结束，而该函数可利用源自非参加者的数据得到估计。当数据来自设定良好且已执行的社会实验时，计算的简单性应该被认为是对照现实社会实验的背景，社会实验受限于其他一些问题，诸如随机化偏倚(**randomization bias**)以及替代偏倚(**substitution bias**)(第 3 章曾讨论过)。

假定对于已处理参加者来说，其结果方程为：

$$y_1 = \text{E}[y_1 \mid \mathbf{x}] + u_1 \tag{25.21}$$

$$\quad = \mu_1(\mathbf{x}) + u_1 \tag{25.22}$$

而对于非参加者来说，其方程为：

$$y_0 = \text{E}[y_0 \mid \mathbf{x}] + u_0 \tag{25.23}$$

$$\quad = \mu_0(\mathbf{x}) + u_0 \tag{25.24}$$

注意到，这种设定意义下具有(类似于 16.7 节已讨论的罗伊模型)转换回归形式，即已处理组与未处理组具有不同条件均值函数 $\mu_1(\mathbf{x})$与 $\mu_0(\mathbf{x})$，这两个函数可用比纯

线性模型所必需的记号更为一般的形式写出来。我们假定，$E[u_1|\mathbf{x}]=E[u_0|\mathbf{x}]=0$，尽管 $E[u_1|\mathbf{x},D]$ 与 $E[u_0|\mathbf{x},D]$ 不一定等于 0。

一种普遍却有约束性的设定为：

$$\mu_1(\mathbf{x})=\mu_0(\mathbf{x})+\alpha D \tag{25.25}$$

其中，已处理组含有附加截距成分 α，但回归元的斜率系数并没有受到处理影响。

观测到的结果可写成：

$$y=Dy_1+(1-D)y_0 \tag{25.26}$$

将上述这些式子组合起来，得到：

$$\begin{aligned} y &=D(\mu_1(\mathbf{x})+u_1)+(1-D)(\mu_0(x)+u_0) \\ &=\mu_0(\mathbf{x})+D(\mu_1(\mathbf{x})-\mu_0(\mathbf{x})+u_1-u_0)+u_0 \end{aligned} \tag{25.27}$$

因为 $D=1$ 或者 0，所以回归的第二项“转换”成开或关。式(25.27)中的第二项测量了参与利益；其第一个成分 $\mu_1(\mathbf{x})-\mu_0(\mathbf{x})$ 测算了具有特征 $\mathbf{x}$ 的参加者平均增益，而第二个成分 (u_1-u_0) 测算的是特定个体的利益。第二个成分可以被参加者观测到，但不能被研究者观测到。

对于一般情况及特殊化的式(25.25)来说，ATE 与 ATET 的表达式已由表 25.2 给出。

表 25.2　处理效应测量：ATE 与 ATET

测量	处理效应	特殊情况(25.25)
给定 $\mathbf{x}$ 时 ATE	$E[\Delta\|\mathbf{x}]=\mu_1(\mathbf{x})-\mu_0(\mathbf{x})$	$E[\Delta\|\mathbf{x}]=\alpha$
含有 $\mathbf{x}$ 及选择效应的 ATET	$E[\Delta\|\mathbf{x},D=1]$ $=\mu_1(\mathbf{x})-\mu_0(\mathbf{x})+E[u_1-u_0\|\mathbf{x},D=1]$	$E[\Delta\|\mathbf{x},D=1]$ $=\alpha+E[u_1-u_0\|\mathbf{x},D=1]$
含有 $\mathbf{x}$ 个体的额外利益	$E[u_1-u_0\|\mathbf{x},D=1]$	$E[u_1-u_0\|\mathbf{x},D=1]$
平均选择偏倚	$E[u_0\|\mathbf{x},D=1]-E[u_0\|\mathbf{x},D=0]$	$E[u_0\|\mathbf{x},D=1]-E[u_0\|\mathbf{x},D=0]$

平均选择偏倚是处于基本状态下项目参加者与非参加者之间的差异。这种效应不能归因于项目。一种特殊情况是，$E[u_1-u_0|\mathbf{x},D=1]=0$，若利益不存在不可观测成分，或者 u_1-u_0 的最佳个体估计是 0，则会出现这种情况。

当结果方程中的处理变量与误差相关时，就产生了选择偏倚。这个相关性是由不正确省略了可观测变量而引起的，而省略掉的可观测变量会部分决定 D 与 y_0。于是，回归误差的省略变量成分将是与 D 相关的，这正是基于可观测成分选择(**selection on observables**)的情况。另一个根源包括既决定 D 又决定 y 的不可观测因素。这是基于不可观测成分选择(**selection on unobservables**)的情况。条件独立性假设本质上会剔除掉由省略变量而引起的混淆。

25.3.3　*对可观测因素的选择*

在观测数据中，基于可观测成分的选择问题可通过利用回归方法与匹配方法

来解决。本章后面几节将详细阐述这些方法。在这样做之前，注意到，16.4 节的两部分模型是一个例子，而在本节我们讨论第二种简单方法。

控制函数估计量(**control function estimator**)是受到决定 D 的可观测变量集合可能与结果相关的可能性而提出的。具体起见，考察结果方程为

$$y_i=\mathbf{x}_i'\boldsymbol{\beta}+\alpha D_i+u_i \tag{25.28}$$

的一种特殊情况，而误差是使得

$$\mathrm{E}[u_i|\mathbf{x}_i,D_i]=\mathrm{E}[u_i|\mathbf{x}_i,D_i,\mathbf{z}_i]$$

在基于可观测成分选择的情况下，有 $\mathrm{E}[u_i|\mathbf{z}_i]\neq 0$。让我们写成：

$$\mathrm{E}[y_i|\mathbf{x}_i,D_i,\mathbf{z}_i]=\mathbf{x}_i'\boldsymbol{\beta}+\alpha D_i+\mathrm{E}[u_i|\mathbf{x}_i,\mathbf{z}_i] \tag{25.29}$$

这就引发了使用建立在方程 OLS/GLS 估计基础上的控制函数估计量。其基本思想是，对可能与 u_i 相关的所有可观测变量引入结果方程，然后通过最小二乘法估计得到的方程。特别地：

$$y_i=\mathbf{C}_i'\boldsymbol{\delta}+\alpha D_i+\{u_i-\mathrm{E}[u_i|D_i,\mathbf{C}_i]\} \tag{25.30}$$

其中，$\mathbf{C}_i$ 包括了所有被 $\mathbf{x}$ 或 $\mathbf{z}$ 所包含的变量。回归中 $\mathbf{z}$ 的存在会剔除 u 与 $\mathbf{z}$ 之间可能的相关性。注意到，若存在基于不可观测成分的选择，它是由既影响 D 又影响 u 的共同不可观测因素引起的，则仍然有潜在的识别问题。

这种估计量被赫克曼和霍茨(Heckman and Hotz, 1989)使用，他们还提出了基本控制函数估计量的一系列变形。

25.3.4 *基于不可观测成分选择*

现在，考察处理参与决策为内生的一种特殊线性情况。这是具有“内生虚拟变量”类型的十分著名的模型。当以观测数据进行研究时，该模型在实证上是非常重要的，因为在这种情况下，存在几种原因放弃约束性假设 $y_0, y_1 \perp D|\mathbf{x}$ 或 $\mathrm{E}[u|\mathbf{x},D]=0$。条件独立性假设失效蕴含着，简单最小二乘法回归不能识别 ATE，从而应致力于一种可供选择的识别策略。

我们将要讨论的识别策略的基本要素为其他选择模型所共有。该方法包括相当强的识别假设，并且是完全参数的。在考虑的特殊情况下，其设定类似于罗伊模型。对结果方程条件的均值采用线性形式。通过添加关于 D_i 的参与(二值)决策方程，可完成此模型。于是：

$$\begin{aligned}y_{1i}&=\mathbf{x}_i'\boldsymbol{\beta}_1+u_{1i}\\ y_{0i}&=\mathbf{x}_i'\boldsymbol{\beta}_0+u_{0i}\\ D_i^*&=\mathbf{z}_i'\boldsymbol{\gamma}+\varepsilon_i\end{aligned} \tag{25.31}$$

其中，D_i^* 表示潜变量，使得：

$$D_i=\begin{cases}1, & \text{当且仅当 } D_i^*>0\\ 0, & \text{当且仅当 } D_i^*\leqslant 0\end{cases} \tag{25.32}$$

而且，假定 $\mathrm{E}[u_1|\mathbf{x},\mathbf{z}]=\mathrm{E}[u_0|\mathbf{x},\mathbf{z}]=0$。

变量 $\mathbf{z}$ 可能与 $\mathbf{x}$ 交叠，但要假定 $\mathbf{z}$ 至少一个成分是唯一的，该成分记为 z_1，并是 D 的非平凡行列式。也就是说，D 至少存在变异的一个独立根源。因此，我们称 z_1 为工具变量，它与内生变量 D 相关，而与结果 y_1 和 y_0 不相关，只是 D 除外。

然后，假定三元组$(u_{1i},u_{0i},\varepsilon_i)$服从联合多变量正态分布，其均值为 0，并且协方差矩阵 $\mathbf{\Sigma}$ 为：

$$\mathbf{\Sigma}=\begin{bmatrix}\sigma_{11} & \sigma_{10} & \sigma_{1\varepsilon}\\ \sigma_{10} & \sigma_{00} & \sigma_{0\varepsilon}\\ \sigma_{1\varepsilon} & \sigma_{0\varepsilon} & 1\end{bmatrix} \tag{25.33}$$

非零协方差矩阵系数 $\sigma_{1\varepsilon}$ 与 $\sigma_{0\varepsilon}$ 反映处理变量的内生性。协方差参数 σ_{10} 反映结果之间的协方差。因为我们永远不能在两个状态下观测到同一个体，所以不能识别这一系数，而通常令其为 0。为了识别，对 ε 的方差限制成 1。

已知这类完全参数设定，通过极大似然法或两步半参数方法对此模型进行估计。这些问题的大多数已经在第 16 章讨论过。将估计问题放在一旁，考察处理影响的测量。

参与的益处或 ATET 可由

$$y_{1i}-\mathrm{E}[y_{0i}|D_i=1]=y_{1i}-\mathbf{x}_i'\boldsymbol{\beta}_0+\sigma_{0\varepsilon}\frac{\phi(\mathbf{z}_i'\boldsymbol{\gamma})}{(1-\Phi(\mathbf{z}_i'\boldsymbol{\gamma}))} \tag{25.34}$$

给出，它还可写成：

$$\mathrm{E}[y_{1i}|D_i=1]-\mathrm{E}[y_{0i}|D_i=1]=\mathbf{x}_i'(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)+(\sigma_{0\varepsilon}-\sigma_{1\varepsilon})\frac{\phi(\mathbf{z}_i'\boldsymbol{\gamma})}{\Phi(\mathbf{z}_i'\boldsymbol{\gamma})} \tag{25.35}$$

其中，$(\sigma_{0\varepsilon}-\sigma_{1\varepsilon})\phi(\mathbf{z}_i'\boldsymbol{\gamma})/\Phi(\mathbf{z}_i'\boldsymbol{\gamma})$表示选择效应(**selection effect**)，参见 16.7.1 节。

在 $\mathbf{x}_i'\boldsymbol{\beta}_0=\mathbf{x}_i'\boldsymbol{\beta}_1$ 且处理虚拟变量以线性方式含有 α 系数进入方程的特殊情况下，项目的平均影响由

$$\mathrm{E}[y_i|D_i=1]-\mathrm{E}[y_i|D_i=0]=\alpha+\text{选择项} \tag{25.36}$$

给出。

在某些样本情形下，这种识别策略或许有点脆弱。例如，已处理组与未处理组可以截然不同，多变量正态性假设可能显得不合适，或者识别工具变量 z_1 与结果方程的误差可能是弱相关的或相关的。

这些考虑激发了使用本章阐述过的可供选择的估计方法。这些估计量通常假定仅仅基于可观测成分选择，此外尽管当选择是基于不可观测成分进行时，25.7 节将阐述利用Ⅳ方法。

25.4 匹配估计量与倾向得分估计量

在观测研究中，由定义知，不存在实验控制。因此，不存在作为已处理组与未处理组之间平均差异计算 ATE 的直接对应形式。换句话说，反事实是不可识别的。作为一种替代，我们从一系列潜在的比较单元中获得数据，比较单元不一定从同一总体中抽取作为已处理单元，但对此而言，可观测特征 $\mathbf{x}$ 会匹配到已处理单元

的那些特征已经达到某种选择的密切程度。

在没有处理的情况下，未处理匹配组的平均结果可识别出已处理组的平均对应结果。这种方法通过假定选择与以 $\mathbf{x}$ 为条件的未处理的结果是不相关的来解决评估问题。为了实施此方法，有必要定义出匹配准则。

25.4.1 *处理效应假设*

当对处理进行选择仅仅是基于可观测成分实施时，处理效应的匹配估计量是有用的。此外，假定交叠（或支撑）条件[**overleap** (or **support**) **condition**](25.6)可以应用，它意味着对于每一个 $\mathbf{x}$，存在一个正的非参与概率。这一点确保我们对于第一个 $\mathbf{x}$ 来说，都拥有未处理的匹配到已处理观测值。粗略地说，控制总体与已处理总体具有可比较的观测特征。生成好的匹配意指可以确保支撑条件不失效。进一步地，重要条件是，不可观测变量在处理指派与结果确定中不起作用。

回归估计量是利用估计回归函数对缺失潜在结果进行估算。当 $D_i=1$ 时，$y_{0,i}$ 就是利用估计条件回归函数 $\hat{\mu}_0(\mathbf{x}_i)$ 估算的。匹配估计量是利用“最近邻”的结果对缺失值加以估算；而“最近邻”是通过基本某个可观测特征的合适距离来定义的。这是匹配估计量典型地逼近于平均值之差，而估计量的方差是利用平均值之差方差的许多可利用结果估计出。

匹配是一种令人信服且吸引人的方法，如果：(1) 我们能控制变量 $\mathbf{x}$ 的丰富集合；(2) 存在许多潜在控制；(3) ATET 是关注的参数。它还需要“无一般均衡效应”假设，或者稳定单元处理值假设（**stable unit treatment value assumption**，记为 **SUTVA**），这蕴含着处理没有间接地影响到未处理观测值。匹配估计量避开了处理效应以线性方式进入条件均值函数的假设。对每个观测值而言，建立其最近匹配的最初步骤，也将会澄清可比较控制观测值是否是有价值的。与回归方法不同，将范围内结论外推到数据邻域之外犯错的危险很小。

假定处理情况是以所有可观测协变量来进行匹配的。在受约束意义上，已处理组与未处理组之间的所有差异都是可控制的。给定结果 y_{1i} 和 y_{0i}，对于处理组与对照组来说，它们的平均处理效应分别为：

$$\begin{aligned}&\mathrm{E}[y_{1i}\mid D_i=1]-\mathrm{E}[y_{0i}\mid D_i=0]\\&\quad=\mathrm{E}[y_{1i}-y_{0i}\mid D_i=1]+\{\mathrm{E}[y_{0i}\mid D_i=1]-\mathrm{E}[y_{0i}\mid D_i=0]\}\end{aligned}\tag{25.37}$$

第二行的第一项是 ATET，而大括号中第二项是“偏差”项，倘若指派处理与控制是随机进行的，则偏差项将为 0。在此情况下，必须估计 ATET 的所有内容就是对因处理而导致的差异进行平均。

更为现实地讲，数据将牵涉到某些观测到的协变量。假定协变量包括涉及选择进入处理组的决定因素的变量。如果已处理组与未处理组在协变量的每个组合上都加以匹配，那么对于每个处理情况及每个 $\mathbf{x}_i$ 来说，处理差异能很容易计算出来。对所有已处理个体与所有 $\mathbf{x}_i$ 上的差异进行平均，该值就测算了平均处理效应。正式地讲，在此情况下[参见安格里斯特和克鲁格(Angrist and Krueger, 2000，第 1 316 页)]，处理对已处理的效应由：

$$\begin{aligned} \mathrm{E}[y_{1i}-y_{0i}\mid D_i=1] &= \mathrm{E}[\{\mathrm{E}[y_{1i}\mid \mathbf{x}_i, D_i=1]-\mathrm{E}[y_{0i}\mid \mathbf{x}_i, D_i=0]\}\mid D_i=1] \\ &= \mathrm{E}[\Delta_{\mathbf{x}}\mid D_i=1] \end{aligned} \tag{25.38}$$

给出，其中，$\Delta_{\mathbf{x}}=\mathrm{E}[y_{1i}\mid \mathbf{x}_i, D_i=1]-\mathrm{E}[y_{0i}\mid \mathbf{x}_i, D_i=0]$。

当变量是 $\mathbf{x}$ 离散的时候，匹配估计量可被定义成加权和：

$$\mathrm{E}[y_{1i}-y_{0i}\mid D_i=1]=\sum_{\mathbf{x}}\Delta_{\mathbf{x}}\Pr[x_i=\mathbf{x}\mid D_i=1] \tag{25.39}$$

其中，$\Pr[\mathbf{x}_i=\mathbf{x}\mid D_i=1]$表示关于 $\mathbf{x}_i$ 的概率质量。安格里斯特和克鲁格（Angrist and Krueger，2000）曾经讨论过关于该估计量的多方面内容。

25.4.2 *准确匹配*

方法是要依据其可观测特征 $\mathbf{x}$，在已处理个体与未处理个体之间进行匹配。

当协变量向量是离散的，而且样本包含 $\mathbf{x}_i$ 的每个不同值上的众多观测值时，准确匹配（**exact matching**）是可行的。

如果协变量向量具有高维数，或者在某些协变量之间连续变差得以表示，那么在已处理组与未处理组之间准确匹配就会变得不切实际。这一问题激发了非准确匹配（**inexact matching**）方法。非准确匹配是通过利用通常纯量 $f(\mathbf{x})$，将 $\mathbf{x}$ 映射到较低维测度的、连续的或离散的，这里 $f(\mathbf{x})$构成了匹配的基础。

25.4.3 *倾向得分*

倾向得分方法[罗森鲍姆和鲁宾（Rosenbaum and Rubin，1983）]是一种流行的非准确匹配。它是针对倾向得分进行匹配，而不是对回归元进行匹配。这里，准确匹配是不可能的，因此比较单元是那些倾向得分充分接近于处理单元的单元。

倾向得分（**propensity score**），即给定时 $\mathbf{x}$ 接收处理的条件概率，记为 $p(\mathbf{x})$，是由罗森鲍姆和鲁宾（Rosenbaum and Rubin，1983）提出作为匹配测量。正如25.2.5 节阐述的，如果数据被证明针对 $\mathbf{x}$ 匹配正确，那么建立在倾向得分基础上的匹配同样可以被证明是正确的。

倾向得分通常利用参数模型，诸如 logit 或 probit 来进行估计，但在原则上，也能利用非参数方法加以估计。

利用倾向得分匹配

在倾向得分方法中，人们可通过控制协变量的特殊函数，尤其是处理的条件概率 $\Pr[D_i=1\mid \mathbf{x}_i]$来控制协变量。也就是说，匹配是针对倾向得分的。这可以很容易地借助于（例如）logit 回归加以计算。此外，人们还能借助于协变量向量包括滞后变量来控制滞后变量。如果选择偏倚可通过控制 $\mathbf{x}_i$ 而得以剔除，那么它也可通过控制倾向得分来剔除。以倾向得分为条件常常比以大维数向量 $\mathbf{x}$ 为条件简单。德赫贾和沃赫拜（Dehejia and Wahba，1998）提供了建立在以前曾由拉隆德使用过的数据基础上的实证说明。

实施问题

倾向得分方法需要好模型来生成得分。我们关注的内容是一致地估计参与概率，而不是倾向得分函数中的参数估计。对于倾向得分来说，一个较好的统计拟合

可能是由灵活的参数模型或非参数模型引起的。

基于 $p(\mathbf{x}_i)$实施匹配的三个有关问题是：(1) 匹配是放回的还是不放回的；(2) 用于比较集合的单元个数；(3) 对匹配方法的选择。

不放回匹配意指，比较组中的任何一个观测值仅仅只与一个已处理观测值进行匹配，这样做是最接近匹配；而放回匹配意指，存在多重匹配。如果匹配不放回，比较集合的最小性意指，匹配根据 $p(\mathbf{x})$不可能是非常接近的，这样将会使估计量的偏差增大。

在比较集合时，选取案例数目问题会牵扯到在偏倚与方差之间的权衡。通过利用对已处理案例的单一最接近匹配，人们可减少偏倚，但通过包括更多匹配控制个体，就使方差减少而偏倚增大，对于已处理观测值来说，额外的观测值都是较差的匹配。部分解决方法是依据已处理观测值的 $p(\mathbf{x})$半径使用预先定义的邻域，同时去掉位于这个邻域之外的匹配。换句话说，人们仅仅使用较好的匹配。这是所谓的“测径匹配”(**caliper matching**)。

赫克曼等人(Heckman et al., 1997, 1998)将从职业培训协作法(Job Training Partnership Act，记为JTPA)获得的实验数据与源自三个来源的比较组的样本结合起来，研究匹配估计量的效果。数据质量在利用匹配方法对处理效应进行稳健估计中起着重要作用。当数据来源及定义对于已处理组与未处理组而言都是可比较的时候，当已处理个体与未处理个体均来自同一个劳动力市场时，并且当倾向得分都可利用回归元的丰富集合进行建模时，结果将是最好的。

结果对选取方法的敏感性问题并不易做出简单而直接的回答。其结论会随着各种不同样本而变化，依赖于已处理与未处理观测值之间的交叠程度。倘若两个组在倾向得分上存在大量交叠的意义是相似的，同时比较组又大，则很容易找到匹配，并且放回匹配(**matching with replacement**)将是可行的。如果比较组是小的且根本不同于已处理组，那么人们可以用完满意的匹配，同时不能使用所有的已处理样本，若匹配是不放回的，尤其可能就是这种情况。

德赫贾和沃赫拜(Dehejia and Wahba, 2002)利用国家支持工作项目(**National Supported Work Program**)数据提供了一种有启发性的例子。我们将在25.8节利用德赫贾和沃赫拜数据集对实施问题加以仔细考察阐明。

25.4.4 *测量处理效应*

把含有特征的已处理案例 i 的比较组作为集合 $A_j(\mathbf{x})=\{j\,|\,\mathbf{x}_j\in c(\mathbf{x}_i)\}$，其中，$c(\mathbf{x}_i)$表示 $\mathbf{x}_i$ 的特征邻域。设 N_C 表示比较组中案例个数，而设 $w(i,j)$表示在与第 i 个已处理案例比较时对第 j 个案例给予的权数，$\sum_j w(i,j)=1$。匹配ATET估计量的一般公式(**general formula**)是：

$$\Delta^M=\frac{1}{N_T}\sum_{i\in\{D=1\}}\left[y_{1,i}-\sum_j w(i,j)y_{0,j}\right] \tag{25.40}$$

其中，$0<w(i,j)\leqslant 1$，$\{D=1\}$表示已处理个体的集合，j 表示已匹配比较单元集合的元素。各种不同的匹配估计量，可通过变动 $w(i,j)$的选取来生成。

匹配方法

简单匹配是把单元(**cells**)与完全相同的离散进行比较:

$$\Delta^M = \sum_k w_k(\bar{y}_{1,k} - \bar{y}_{0,k}) \tag{25.41}$$

其中,$\bar{y}_1$ 表示已处理的平均结果,$\bar{y}_0$ 表示未处理的平均结果,而 $\bar{w}_k$ 表示第 k 个单元的权数(也就是说,在单元 k 部分的观测值)。

一个特定例子[德赫贾和沃赫拜(Dehejia and Wahba, 2002)]是:

$$\frac{1}{N_T}\sum_i \left(y_i - \frac{1}{N_{C,i}}\sum_{j\in\{D=0\}} y_j\right) \tag{25.42}$$

其中,N_T 表示已处理组($D=1$)的个数,而 $N_{C,i}$表示对应于第 i 个观测值的比较组个数。

对于每个已处理个体来说,最近邻匹配(**nearest-neighbor matching**)方法是选择集合 $A_i(\mathbf{x})=\{j\mid \min_j \parallel \mathbf{x}_i-\mathbf{x}_j \parallel\}$,其中,$\parallel\ \parallel$ 表示向量之间的欧几里得距离(Eudidean distance)。如果当 $j\in A_i(\mathbf{x})$时,式(25.40)中 $w(i,j)=1$,否则为 0,那么这种设定仅仅使用了一个案例来构建对已处理情况的比较组。

另一种估计量是由核匹配(**kernel matching**)生成的,这里有:

$$w(i,j)=\frac{K(\mathbf{x}_j-\mathbf{x}_i)}{\sum_{j=1}^{N_{C,i}} K(\mathbf{x}_j-\mathbf{x}_i)}$$

其中,K 表示已在 9.3 节讨论的核。

这些方法分享了在估计 ATET 时避免关于结果方程的函数形式假设的优点,同时在 $\mathbf{x}$ 的特定值上对其进行估计。这些方法具有下述缺点:如果 $\mathbf{x}$ 具有高维,那么匹配个数变得非常少。在这种情况下,基于纯量值距离的匹配是引人注目的。前面已讨论的倾向得分匹配(**propensity score matching**)正是此类方法。

最近邻匹配与核匹配还能用倾向得分加以定义。例如,对最近邻匹配来说,可将其匹配定义为 $A_i(p(\mathbf{x}))=\{p_j\mid \min_j \parallel p_i-p_j \parallel\}$。

分层匹配或区间匹配(**stratification or interval matching**)是基于对区间上倾向得分变化范围加以分割,使每个区间内的已处理单元与对照单元就平均水平而言拥有相同倾向得分。人们能使用由计算倾向得分的算法识别的相同块。然后,计算已处理组与对照组的平均结果之间的差异。ATET 是这些差异的加权平均,其权数可利用已处理单位在各种不同块之间的分布来确定。这种方法的缺点之一是,它丢弃了块中缺乏的已处理单位或对照单元中的观测值。

用 b 表示在倾向得分区间上定义的块。于是,将第 b 块内的处理效应定义成:

$$\text{ATET}_b^S = (N_b^T)^{-1}\sum_{i\in I(b)} Y_{1i} - (N_b^C)^{-1}\sum_{j\in I(b)} Y_{0j}$$

其中,$I(b)$表示块中的单位集合,N_b^T 表示第 b 块内已处理单位的个数,而 N_b^C 表示第 b 块内控制单位的个数。然后,将基于层的处理效应定义成:

$$\text{ATET}^S = \sum_{b=1}^{B} \text{ATET}_b^s \times \left[\sum_{i\in I(b)} D_i \Big/ \sum D_i\right] \tag{25.43}$$

其中，括号内的项表示对应于已处理单元的部分所给出的每一块的权数，这里，B 表示总块数。

在半径匹配（**radius matching**）中，集合 $A_i(p(\mathbf{x}))=\{p_j \mid \| p_i-p_j \| <r\}$ 是建立在倾向得分的基础上。这意味着，含有估计倾向得分落入半径之间的所有对照案例都匹配到第 i 个已处理案例。

一旦假定交叠条件 $0<p(\mathbf{x})<1$，就能用 $p(\mathbf{x})$ 表示 ATE 与 ATET。两个重要结论是：

$$\text{ATE}=\mathrm{E}\left[\frac{(D-p(\mathbf{x}))y}{p(\mathbf{x})(1-p(\mathbf{x}))}\right] \tag{25.44}$$

$$\text{ATET}=\mathrm{E}\left[\frac{(D-p(\mathbf{x}))y}{\Pr[D=1](1-p(\mathbf{x}))}\right] \tag{25.45}$$

第二个结论归功于德赫贾（Dehejia，1997）。

这些结论的推导如下：

$$\begin{aligned} y &=(1-D)y_0+Dy_1 \\ &=y_0+D(y_1-y_0), \\ (D-p(\mathbf{x}))y &=(D-p(\mathbf{x}))(y_0+D(y_1-y_0)) \\ &=Dy_1-p(\mathbf{x})y_0-Dp(\mathbf{x})y_1+Dp(\mathbf{x})y_0 \\ &=Dy_1-p(\mathbf{x})(1-D)y_0-Dp(\mathbf{x})y_1 \end{aligned} \tag{25.46}$$

其次，取期望，并注意到，$\mathrm{E}[\mathrm{D}|\mathbf{x}=p(\mathbf{x})]$，我们得到：

$$\begin{aligned} \mathrm{E}[(\mathrm{D}-p(\mathbf{x}))y|\mathbf{x}] &=p(\mathbf{x})\mathrm{E}[y_1]-p(\mathbf{x})(1-p(\mathbf{x}))\mathrm{E}[y_0]-p^2(\mathbf{x})\mathrm{E}[y_1] \\ &=p(\mathbf{x})\mathrm{E}[y_1-p(\mathbf{x})y_1]-p(\mathbf{x})(1-p(\mathbf{x}))\mathrm{E}[y_0] \\ &=p(\mathbf{x})(1-p(\mathbf{x}))\mathrm{E}[y_1-y_0] \end{aligned} \tag{25.47}$$

由此可得：

$$\text{ATE}=\mathrm{E}[y_1-y_0]=\mathrm{E}\left[\frac{(D-p(\mathbf{x}))y}{p(\mathbf{x})(1-p(\mathbf{x}))}\right]$$

为了推导德赫贾结论，有：

$$\begin{aligned} \mathrm{E}\left[\frac{(D-p(\mathbf{x}))y}{1-p(\mathbf{x})}\right] &=\mathrm{E}[p(\mathbf{x})\mathrm{E}[\mu_1(\mathbf{x})-\mu_0(\mathbf{x})] \\ &=\mathrm{E}[D(y_1-y_0)] \\ &=\mathrm{E}[D(y_1-y_0)|D=1]\Pr[D=1] \end{aligned} \tag{25.48}$$

其中，第一行是由式(25.47)得到，第二行是由条件独立性假设得出的，而最后行含有期望的表达式作为边缘期望与条件期望的积，这蕴含着：

$$\text{ATET}=\frac{\mathrm{E}[D(y_1-y_0)]}{\Pr[D=1]}$$

利用式(25.44)与式(25.45)，建立在容量为 N 的样本基础上的一致估计量是：

$$\widehat{\mathrm{ATE}}=\frac{1}{N}\sum_{i=1}^{N}\left[\frac{(D_i-\hat{p}(\mathbf{x}_i))y_i}{\hat{p}(\mathbf{x}_i)(1-\hat{p}(\mathbf{x}_i))}\right] \tag{25.49}$$

$$\widehat{\mathrm{ATET}}=\left(\frac{1}{N}\sum_{i=1}^{N}D_i\right)^{-1}\sum_{i=1}^{N}\left[\frac{1}{N}\frac{(D_i-\hat{p}(\mathbf{x}_i))y_i}{(1-\hat{p}(\mathbf{x}_i))}\right] \tag{25.50}$$

其中，$(N^{-1}\sum_{i=1}^{N}D_i)$表示 $\Pr[D=1]$的一致估计量。

25.4.5 基于 x 与 p(x)的 ATET 方差

在 25.2 节给出的识别性假设下，$\hat{\Delta}_{\mathbf{x}}$ 与 $\hat{\Delta}_{p(\mathbf{x})}$ 可被定义成：

$$\begin{aligned}\hat{\Delta}_{\mathbf{x}}&=\frac{1}{N_T}\sum[y_{1i}-\hat{\mathrm{E}}[y_0\mid D=0,\mathbf{x}=\mathbf{x}_i]]\\&=\frac{1}{N_T}\sum_{i\in\{D=1\}}\left[y_{1i}-\sum_{j\in A_i(\mathbf{x})}w_{ij}y_{0,j}\right]\end{aligned}$$

而：

$$\begin{aligned}\hat{\Delta}_{p(\mathbf{x})}&=\frac{1}{N_T}\sum[y_{1i}-\hat{\mathrm{E}}[y_0\mid D=0,p(\mathbf{x})=p(\mathbf{x}_i)]]\\&=\frac{1}{N_T}\sum_{i\in\{D=1\}}\left[y_{1i}-\sum_{j\in A_ip(\mathbf{x})}w_{ij}y_{0,j}\right]\end{aligned}$$

其中，i 表示已处理组的下标，$w_{ij}=1/N_{C,i}$表示关于第 i 个已处理组的比较组中案例个数。这两个均是 ATET，$\mathrm{E}[y_1-y_0\mid D=1,\mathbf{x}]$的一致估计量，第一个建立在 $\mathbf{x}$ 基础上，而第二个则建立在 $p(\mathbf{x})$基础上。一个实际问题是，就有效性而论，依据倾向得分对差进行调整是否比利用 $\mathbf{x}$ 对差进行调整更好些。哈恩(Hahn, 1998)、赫克曼等人(Heckman et al., 1998)以及其他一些人曾证明，即使我们假定 $p(\mathbf{x}_i)$是已知的，但根据估计量渐近方差来看，对这两个估计量进行排序就会含糊不清，在观测研究时将不会出现此种情况。

对上述两种情况的渐近方差，可写成如下形式：

$$\mathrm{V}[\hat{\Delta}_{\mathbf{x}}]=\mathrm{E}[\mathrm{V}[y_1\mid D=1,\mathbf{x}]\mid D=1]+\mathrm{V}[\mathrm{E}[y_1-y_0\mid D=1,\mathbf{x}]\mid D=1]\ ,$$

$$\mathrm{V}[\hat{\Delta}_{p(\mathbf{x})}]=\mathrm{E}[\mathrm{V}[y_1\mid D=1,p(\mathbf{x})]\mid D=1]+\mathrm{V}[\mathrm{E}[y_1-y_0\mid D=1,p(\mathbf{x})]\mid D=1]$$

其中，我们运用了由 A.8 节给出的方差分解结论。通常，与 $p(\mathbf{x})$相比，$\mathbf{x}$ 是一个更好的预测式，这蕴含着：

$$\mathrm{E}[\mathrm{V}[y_1\mid D=1,\mathbf{x}]\mid D=1]\leqslant\mathrm{E}[\mathrm{V}[y_1\mid D=1,p(\mathbf{x})]\mid D=1]$$

$$\mathrm{V}[\mathrm{E}[y_1-y_0\mid D=1,\mathbf{x}]\mid D=1]\geqslant\mathrm{V}[\mathrm{E}[y_1-y_0\mid D=1,p(\mathbf{x})]\mid D=1]$$

因为以 $\mathbf{x}$ 为条件所损失的信息比以 $p(\mathbf{x})$为条件的要少一些，它是 $\mathbf{x}$ 的一个特定函数。因而，第二个比较有利于倾向得分方法，而第一个比较有利于对 $\mathbf{x}$ 的使用，而不是对 $p(\mathbf{x})$的使用。

实施 ATET 计算的一个有用的实践指南与计算机程序是由贝克尔和市野(Becker and Ichino, 2002)提供的。

25.5 差异中差分估计量

第 2 章与第 3 章曾讨论过自然实验(**natural experiment**)或准实验(**quasi-experiment**)的设置背景,其中,处理变量经受了能被看成处理变量中外生变动的变化。已处理组能与未处理的比较组加以比较。

在一些情况下,人们拥有实验前后的已处理组与比较(控制)组方面的数据。然后,对于第 i 个已处理案例来说,其结果变化可由$[y_{ia}-y_{ib}|D_{ia}=1]$测算,而关于未处理组的结果变化可由$[y_{ia}-y_{ib}|D_{ia}=0]$测算。于是,差异中差分测算$[y_{ia}-y_{ib}|D_{ia}=1]-[y_{ia}-y_{ib}|D_{ia}=0]$,其中,下标 a 与 b 分别表示实验发生“以后”与“之前”,构成了处理效应估计的基础。这一方法已在 3.4.2 节与 22.6 节介绍过。

考虑含有固定效应 ϕ_i 与漂移项 δ_t 的模型,其中,处理前结果与处理后结果分别由

$$y_{it,0}=\phi_i+\delta_t+\varepsilon_{it} \tag{25.51}$$

$$y_{it,1}=y_{it,0}+\alpha \tag{25.52}$$

给出,因此:

$$\begin{aligned} y_{it}&=(1-D_{it})y_{it,0}+D_{it}y_{it,1} \\ &=\phi_i+\delta_t+\alpha D_{it}+\varepsilon_{it} \end{aligned} \tag{25.53}$$

上述式子是关于 $t=a,b$ 的;式(25.51)是没有接受处理组的,而式(25.52)是接受处理组的。一旦利用“之前”与“以后”公式,就得出处理效应:

$$\begin{aligned} \alpha&=\mathrm{E}[y_{ia}-y_{ib}|D_{ia}=1]-\mathrm{E}[y_{ia}-y_{ib}|D_{ia}=0] \\ &=\{\mathrm{E}[y_{ia}|D_{ia}=1]-\mathrm{E}[y_{ia}|D_{ia}=0]\} \\ &\quad-\{\mathrm{E}[y_{ib}|D_{ia}=1]-\mathrm{E}[y_{ib}|D_{ia}=0]\} \end{aligned} \tag{25.54}$$

其中,进行差分步骤去掉了固定效应 α 与漂移 δ_t。

存在一些可供选择的进行差分方法。一种可选择的方法是,通过回归直接控制处理组与对照组之间的处理前结果。例如,用 $\mathbf{x}_i'\boldsymbol{\beta}+\gamma y_{ib}$代替式(25.51)中的 ϕ_i,得出:

$$\begin{aligned} y_{ia,0}&=\mathbf{x}_i'\boldsymbol{\beta}+\gamma y_{ib}+\delta_a+\varepsilon_{ia} \\ y_{ia,1}&=\mathbf{x}_i'\boldsymbol{\beta}+\gamma y_{ib}+\delta_a+\alpha D_{ia}+\varepsilon_{ia} \end{aligned} \tag{25.55}$$

对 α 的估计可通过处理后结果对常值、处理前结果 $\mathbf{x}_i$ 以及 D_{ia} 进行回归构造出来。对作为因果参数 α 进行解释依赖于下述假设:一旦控制 $\mathbf{x}$ 与 y_b 之后,处理效应完全说明了已处理组与对照组之间的处理后差异。固定效应是由线性函数形式给出的,而匹配策略则可以建立在弱假设的基础上。

实际上,前面结论建立在准实验数据基础上。例如,将一个州有某一法律与不同州具有不同法律的人们进行比较,而且使用州效应的控制函数。在实验之前,就要增加新的数据。借助于两个州具有相同漂移项的假设,我们能应用差异中差分方法剔除州效应,否则,就需要控制函数。

25.6 回归非连续设计

有时，对处理效应的识别或者借助于自然实验或者利用在准实验背景下生成的数据使其便利。非连续回归(**regression discontinuity**，记为RD,又称为回归间断点方法)设计是准实验设计的一个例子，其中，接收处理的概率是一个或多个基本变量的非连续函数。这种设计是在管理或组织控制为处理的直接原因情况下而产生的。例如，安格里斯特和拉维(Angrist and Lavy，1999)曾研究了班级大小对学生分数的效应，他们利用在“迈蒙尼德斯规则”(**Maimonides Rule**)作用下生成的数据，即约定当分数达到某个特定门限水平时，班级要分班。范德克劳(Van der Klaauw，2003)估计了提供资金援助对学生决策上大学的效应，利用了由管理规则上非连续所提供的识别信息，管理规则与援助学生的SAT分数以及平均成绩(**grade point average**)有关。这些经济计量应用的先驱是西斯尔思韦特和坎贝尔(Thistlethwaite and Campbell，1960)，他们分析了学生对职业意愿的影响，利用当学生考试分数大于某个分值门限时，才能给予奖金的事实，也可参见特罗布姆(Trochim，1984)。这里的研究遵循范德克劳(Van der Klaauw，2003)的线索。

25.6.1 *非连续处理指派机制*

在RD设计情况下，选择规则方面存在额外信息：众所周知，处理指派机制(至少部分地)依赖于与给定门限或截止分数有关的观测连续变量的值，在这种方式下，相对应的得到已处理(倾向得分)的概率是此变量在截止分数上的非连续函数。图25.1阐明了由RD设计生成的样本。

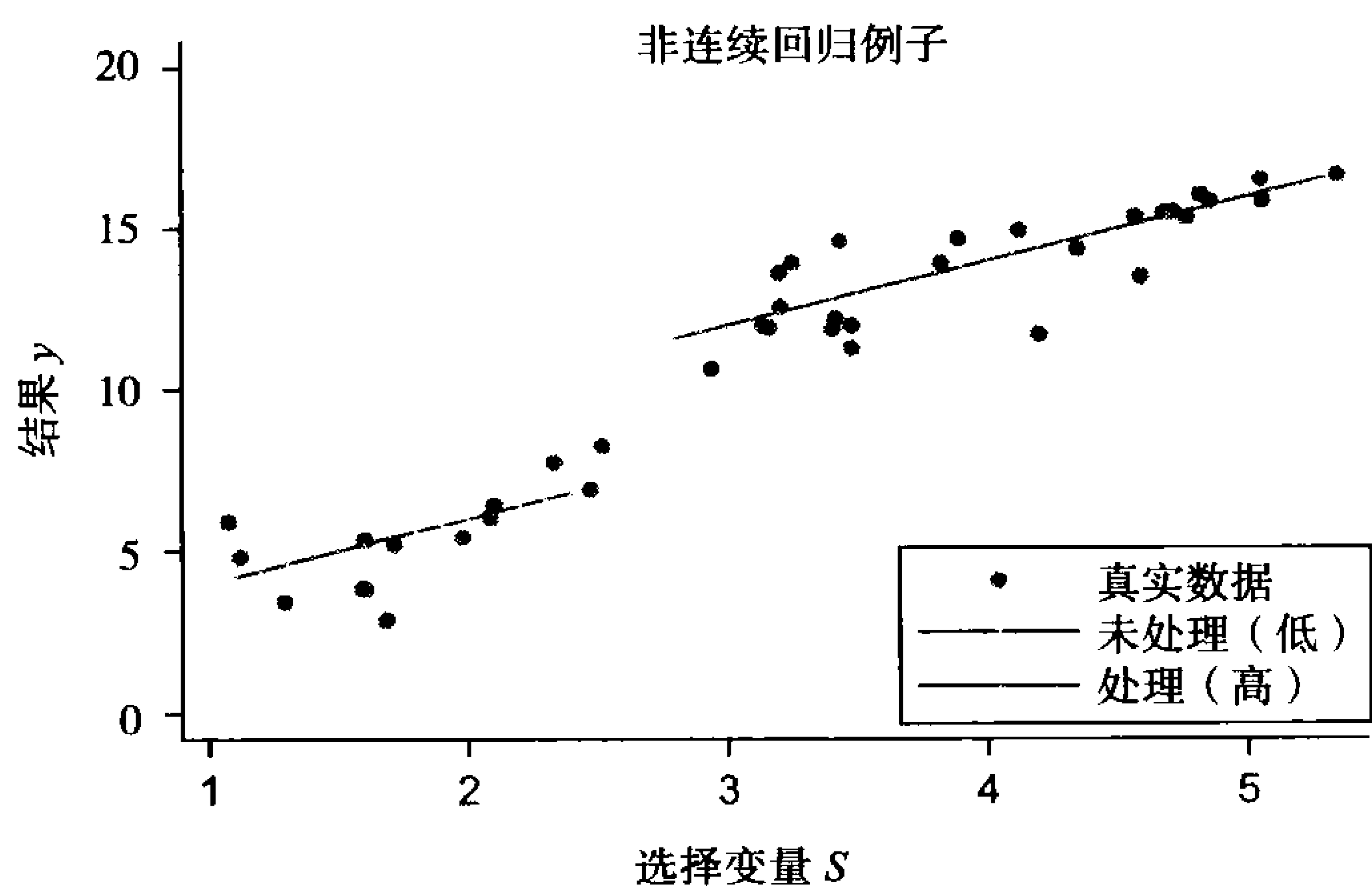


图 25.1 非连续回归设计例子

在最简单RD设计中，所谓的标准RD设计(**sharp RD design**)，是个体唯一地在观测连续测量 S 的基础上被指派到处理组与对照组，其中，S 称为选择或指派变量。落入明显断开 $\bar{S}$ 下面的那些不接收处理，并且构成对照组，而位于断开上面的那些则接收处理($D=1$)。也就是说，处理指派是通过一个已知的且测量的确定性

决策规则而发生：$D_i=\mathbf{1}[S_i\geqslant\bar{S}]$。如图 25.2 所示，标准 RD 设计以实线画出[参见范德克劳(Van der Klaauw，2003)]。

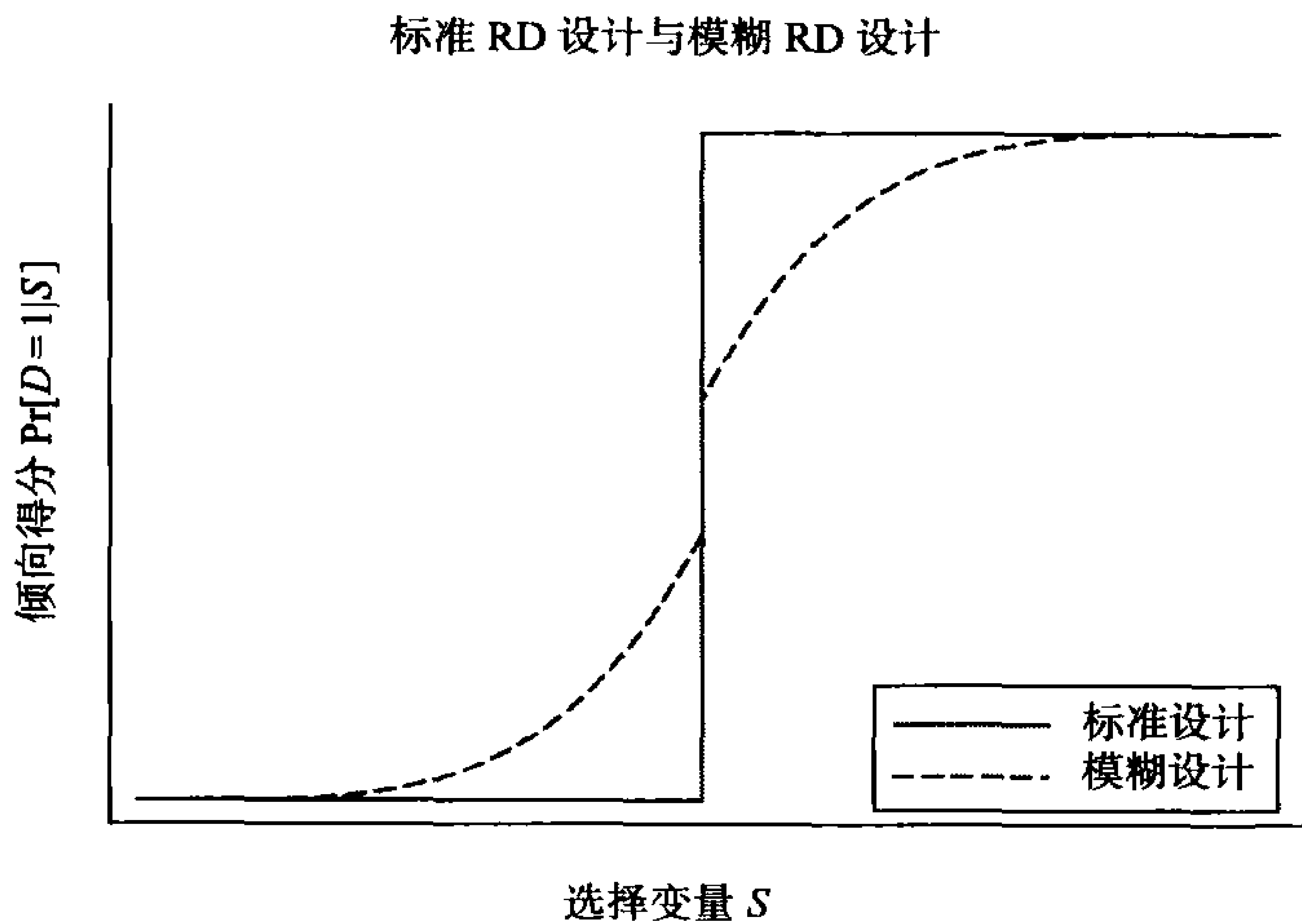


图 25.2 非连续回归设计；标准设计(实线)中的处理指派与模糊设计(虚线)中的处理指派。

在标准 RD 设计中：

$$E[u|D,S]=E[u|S] \tag{25.56}$$

其中，u 表示结果方程中的误差。因为 S 是 D 的唯一系统行列式，所以 S 将捕获 D 到与 u 之间的任何相关性。

对于 $D_i=D(S_i)=\mathbf{1}[S_i\geqslant\bar{S}]$来说，$D_i$ 与 u_i 之间的相依性会导致 OLS 产生的非一致估计量。如同前面提及的，在此情况下，估计处理效应的一种方法是，设定并包括条件均值函数 $E[u|D,S]$ 作为结果方程中的"控制函数"。因而：

$$y_i=\beta+\alpha D_i+k(S_i)+\varepsilon_i \tag{25.57}$$

其中，$\varepsilon_i=y_i-E[y_i\mid D_i,S_i]$。如果 $k(S)$得以正确设定，该回归将会一致地估计出 α。

当 $k(S)$是线性的时候，α 将通过在断开点处两条线性平行回归线之间的距离加以估计，在此情况下，它就等于两个截距之差。若控制函数是线性的，则它是共同处理效应的无偏估计。

在更一般的可变处理效应情况下，D 的系数表示成 $E[\alpha_i|\bar{S}]$或者 25.7.1 节将讨论的局部 LATE，其中，$k(S)$表示 $E[u|S]+(E[\alpha_i|S]-E[\alpha_i|\bar{S}])\mathbf{1}[S\geqslant\bar{S}]$的设定，这里，$\mathbf{1}[S\geqslant\bar{S}]=1$，当括号中的条件得到满足时。对 $k(S)$的不正确设定会导致非一致性，因此，可能就要尝试半参数设定，例如，$k(S)=\sum_{j=1}^{J}\eta_j S^j$，其中，$J$ 可能通过适当方法来决定。

变量 S 可能与结果 y 有关，甚至当两个变量之间并不存在因果联系时，这会自动引起 (y,S) 是相关的。这与可避免此类相依性的随机指派形成了对比。

然而，随机指派除了处理接收之外，会使处理组与对照组在一些方面产生雷同，标准 RD 设计至少在关于 S 值上使处理组与对照组产生差异。这违背了罗森鲍姆和

鲁宾(Rosenbaum and Rubin, 1983)的"强可忽略性"(**strong ignorability**)假设,它还要求交叠条件,$0<\Pr[D=1|S]<1$,而在标准 RD 设计模型中,$\Pr[D=1|S]\in[0,1]$。

25.6.2 *在 RD 设计下的识别与估计*

主要直觉是,在截止点很小邻域内的个体样本将类似于在截止点处的随机实验,因为它们基本上具有相同 S 值。刚好在截止点下面的那些个体,预计非常类似于刚好在截止点上面的那些个体。对刚好位于截止点上面的与刚好位于截止点下面的那些个体平均值 y 进行比较,将产生平均处理效应的估计值。

特别是,如果指派变量自身与以处理状况为条件的结果变量有关时,增大截止点附近的区间将会使处理效应的估计产生偏差。若能对这种关系的函数形式做出假设,则可使用更多观测值,同时从截止点上外推平衡随机化实验所揭示的内容。这种双重外推,连同在截止点附近的"随机化实验"的解释,正是支持非连续性回归分析的主要思想[范德克劳(Van der Klaauw, 2003,第 1 258 页)]。

可以发现,在这种 RD 设计中,有:

$$\lim_{S\downarrow \bar{S}}\mathrm{E}[y|S]-\lim_{S\uparrow \bar{S}}\mathrm{E}[y|S]=\alpha+\lim_{S\downarrow \bar{S}}\mathrm{E}[u|S]-\lim_{S\uparrow \bar{S}}\mathrm{E}[u|S] \tag{25.58}$$

一种更正式假定是,在不存在处理情况下,对在 $\bar{S}$ 附近很小区间里具有相似平均结果的一些个体设定如下:

假设 A1. 条件均值函数 $\mathrm{E}[u|S]$ 在 $\bar{S}$ 处是连续的。

假设 A2. 均值处理效应函数 $\mathrm{E}[\alpha_i|S]$在 $\bar{S}$ 处是右连续的:

$$y_i=\beta+\alpha D_i+k(S_i)+\varepsilon_i \tag{25.59}$$

其中,$\varepsilon_i=y_i-\mathrm{E}[y_i|D_i,S_i]$。于是,式(25.58)中的结果成立。

25.6.3 *模糊 RD 设计*

这里,处理指派依赖于以随机方式所选择的变量。可以知道倾向得分 $\Pr[D=1|S]$之间的关系,在 $\bar{S}$ 处具有非连续性。与截止值有关的错误指派的一个可能后果是模糊 RD,在截止点附近的 S 值既出现在处理组中,又出现在对照组中。否则,指派可建立在被处理管理者观测到而项目评估者观测不到的额外变量基础上。因此,与标准 RD 设计有关,模糊 RD 设计(**fuzzy RD design**)选择既依赖于可观测的又依赖于不可观测的成分。如图 25.2 所示,模糊 RD 设计已用虚线画出。

为了识别在 A1 假设下的处理效应,还要探讨选择规则的非连续性。若 $\mathrm{E}[u|S]$在 $\bar{S}$ 处是连续的,则 $\lim_{S\downarrow \bar{S}}\mathrm{E}[y|S]-\lim_{S\uparrow \bar{S}}\mathrm{E}[y|S]=\alpha[\lim_{S\downarrow \bar{S}}\mathrm{E}[D|S]-\lim_{S\uparrow \bar{S}}\mathrm{E}[D|S]]$。因此,处理效应可通过

$$\frac{\lim_{S\downarrow \bar{S}}\mathrm{E}[y|S]-\lim_{S\uparrow \bar{S}}\mathrm{E}[y|S]}{\lim_{S\downarrow \bar{S}}\mathrm{E}[D|S]-\lim_{S\uparrow \bar{S}}\mathrm{E}[D|S]} \tag{25.60}$$

来识别,其中,分母 $\lim_{S\downarrow \bar{S}}\mathrm{E}[D|S]-\lim_{S\uparrow \bar{S}}\mathrm{E}[D|S]\neq 0$,因为已知 $\mathrm{E}[D|S]$在 $\bar{S}$ 处具有非连续性。

在异方差处理响应(**heterogeneous treatment**)的情况下,我们需要一些额外

假设。

假设 A2*. 平均处理效应函数 $E[\alpha_i|S]$在 $\bar{S}$ 处是连续的。

假设 A3. D_i 与在 $\bar{S}$ 附近以 S 为条件的 α 是独立的：

$$y_i=\beta+\alpha E[D_i|S_i]+k(S_i)+\varepsilon_i \tag{25.61}$$

其中，$\varepsilon_i=y_i-E[y_i|D_i,S_i]$。而 $k(S_i)$表示对 $E[u_i|S_i]$的设定形式。

25.6.4 *两阶段估计量*

当 $Cov[D,u]\neq 0$ 时，OLS 回归将产生有偏倚估计值。不过，下述情形能导致一致估计量。考察：

$$y_i=\beta+\alpha E[D_i|S_i]+k(S_i)+\varepsilon_i \tag{25.62}$$

其中，$\varepsilon_i=y_i-E[y_i|S_i]$，而 $k(S_i)$表示对 $E[u_i|S_i]$的设定形式。

步骤 1： 将模糊 RD 设计的倾向得分设定成：

$$E[D_i|S_i]=f(S_i)+\gamma\, 1[S_i\geqslant\bar{S}] \tag{25.63}$$

其中，$f(S_i)$表示 S 的某个连续函数，它在 $\bar{S}$ 处是连续的。通过对 f 的函数形式加以设定(或者以半参数形式或非参数形式估计 f)能估计出 γ，倾向得分函数具有在 $\bar{S}$ 处的非连续性。

步骤 2： 然后，用 $E[D_i|S_i]=Pr[D_i=1|S_i]$的第一阶段估计值代替 D_i，就可以估计控制增广函数结果方程；此估计值在 S 处是非连续的，但所包括的关于$k(S)$的控制函数在 S 中 $\bar{S}$ 处是连续的。在对 $f(S_i)$与 $k(S_i)$的正确设定下，两阶段方法是一致的。

25.7 工具变量法

在最近几年，工具变量法作为一种 MLE 和其他有说服力的参数方法的选择，得到了强劲的发展与支持[安格里斯特、英伯斯和鲁宾(Angrist, Imbens, and Rubin, 1996)]。在基于不可观测成分选择模型方面，这种识别策略是引人注目的(参见 25.3.4 节)。在许多应用中，这类模型是由连续结果变量的线性方程构成的，而连续结果变量的条件均值及方差结构均是设定的，没有任何额外的分布假设。一种重要情况是，连续结果依赖于回归元 **x** 向量以及表示处理参与新决策的单个内生处理(虚拟)变量(D)。称这一方程为参与或选择方程。在更一般设置背景下，可能拥有受限因变量或离散结果，也可能存在多重处理变量。

下面的讨论与本书中几个地方的Ⅳ估计内容相交叉，同时也与选择模型内容交叉。Ⅳ方法允许我们去发展 ATE 参数的另一种“局部”变形。

25.7.1 *局部 ATE (LAET)*

我们重新考察简单线性公式。结果方程是可观测变量及参与指示变量的线性函数：

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \alpha D_i + u_i \tag{25.64}$$

同时，参与决策依赖于称为工具的单个变量 z：

$$D_i^* = \gamma_0 + \gamma_1 z_i + v_i \tag{25.65}$$

其中，D_i^* 表示潜变量，其可观测部分 D_i 是由

$$D_i = \begin{cases} 0, & \text{当 } D_i^* \leqslant 0 \\ 1, & \text{当 } D_i^* > 0 \end{cases} \tag{25.66}$$

生成的。

存在两个假设：

1. 存在一个变量 z，它出现在关于 D 的方程中，而不出现在关于 $\mathbf{y}$ 的方程中。它可能是连续的或离散的，而在特殊情况下，它是二值的。参与方程中将回归元 $\mathbf{x}$ 排除掉是一种简化。参与方程存在的 z 与由结果方程将其同时排除，称为排除性约束(**exclusion restriction**)。这种模型的结构与第 16 章的选择模型接近。

2. $\mathrm{Cov}[z, v] = \mathrm{Cov}[u, z] = \mathrm{Cov}[x, u] = 0$，以及：

$$\mathrm{Cov}[D, z] \neq 0$$

连同第 1 个假设，此假设蕴含着 y 通过 D 仅仅依赖于 z，而 D 以非一般方式依赖于 z，这正如同前面所强调的。

在这些假设下，式(25.6.4)的Ⅳ估计产生$(\boldsymbol{\beta}, \alpha)$的一致估计值。设 $z' = z + \delta$，$\delta \neq 0$，然后，注意到 $\mathrm{E}[D | \mathbf{x}, D(z)] = \Pr[D(z) = 1]$，同时一旦取期望，我们得到：

$$\mathrm{E}[y | \mathbf{x}, D(z)] = \mathbf{x}'\boldsymbol{\beta} + \alpha \Pr[D(z) = 1]$$
$$\mathrm{E}[y | \mathbf{x}, D(z')] = \mathbf{x}'\boldsymbol{\beta} + \alpha \Pr[D(z') = 1]$$

在它们相减之后，得出：

$$\mathrm{E}[y | \mathbf{x}, z'] - \mathrm{E}[y | \mathbf{x}, z] = \alpha[\Pr[D(z') = 1] - \Pr[D(z) = 1]]$$

关于 α 求解此方程，得到由英伯斯和安格里斯特(Imbens and Angrist, 1994)所分析的局部平均处理效应(**local average treatment effect**, 记为 LATE)：

$$\begin{aligned}\alpha_{\mathrm{LATE}} &= \frac{\mathrm{E}[y | \mathbf{x}, z'] - \mathrm{E}[y | \mathbf{x}, z]}{\Pr[D(z') = 1] - \Pr[D(z) = 1]} \\ &= \frac{\int_{R(\mathbf{x})} [\mathrm{E}[y | \mathbf{x}, z'] - \mathrm{E}[y | \mathbf{x}, z]] dF(\mathbf{x} | \mathbf{x} \in R(\mathbf{x}))}{\int_{R(\mathbf{x})} [\Pr[D(z') = 1] - \Pr[D(z) = 1]] dF(\mathbf{x} | \mathbf{x} \in R(\mathbf{x}))} \\ &= \frac{\mathrm{E}[y | z'] - \mathrm{E}[y | z]}{\Pr[D(z') = 1] - \Pr[D(z) = 1]}\end{aligned} \tag{25.67}$$

其中，第二行涉及对 $\mathbf{x}$ 进行平均，其支集用 $R(\mathbf{x})$ 表示。当 $\Pr[D(z') = 1] - \Pr[D(z) = 1] \neq 0$ 时，这个表达式定义良好。该表达式的样本类似形式是已处理与未处理之间的平均差，被由 z 变动引起的已处理部分变化去除。这个估计量就是Ⅳ估计量。一旦利用Ⅳ估计量的有关渐近正态性结论，就能得到 LATE 参数的置信区间。

可以证明,LATE 中的"局部"合格者是正确的,因为它测算了由参与处理作为 z 变动而引起的遵从者的处理效应。依赖于用于估算处理 z 的特定值以及依赖于所选取的特定工具。"运动者"组可以不是整个已处理总体的代表,更不用说整个总体了。因此,LATE 参数关于由工具变动而引起的大政策变化后果方面没有任何信息,而工具变动有别于那些在过去所观测到的情形。

对于二值工具来说,LATE 与Ⅳ估计是等价的,正如安格里斯特等人所证明的(Angrist et al. ,1996,第 447 页)。如果参与方程中出现不止一个工具,因为存在过度识别约束,所以就每个工具而论,所估计的 LATE 参数通常将是不同的。不过,可以构建一种加权平均。

当处理效应不随个体而变化时,就可应用上述分析。然而,若处理效应是异质性的(**heterogeneous**),则引起的变化存在着潜在的混淆:观测到的变化是起因于 z 的变异,还是由 α 的差异而引起的呢?在异质性下,处理效应的特质成分

$$u_{i,1}=u_{i,0}+D_i(\alpha_i(\mathbf{x}_i)-\alpha(\mathbf{x}_i))$$

是 $\alpha_i(\mathbf{x}_i)-\alpha(\mathbf{x}_i)$的函数,参见式(25.27)。于是,前面的一些假设并不足以确定 ATE 或 ATET。对此困难的求解是加上单调性假设(**monotonicity assumption**)作为另外的识别条件。这本质上表明,工具是以单调方式影响参与的,所以如果平均参与可能由给定 $Z=w$ 而非给定 $Z=z$ 时导出的,那么给定 $Z=z$ 时的那些参与人也必是给定 $Z=w$ 时的参与人。

25.7.2 *与其他测量的关系*

α 的Ⅳ估计量与利用两阶段最小二乘法所估计出的值一样,在两阶段最小二乘法中,首先估计接收处理的概率 $\mathrm{E}[D=1|\mathbf{x},\mathbf{z}]$,然后实施结果与拟合概率的回归;当然假定处理效应是可加的。考察Ⅳ估计量的特殊情况,其中,$\mathbf{x}$ 表示纯量且等于 1,而 $\mathbf{z}$ 表示纯量虚拟变量,它刻画了参与处理的合格性,$z=1$ 蕴含着合格性,而 $z=0$ 蕴含着不合格性。

我们将总体分成四种类型:遵从者(**compliers**)(C)、总是接受者(**always-takers**)(A)、永不接收者(**never-takers**)(N),以及违抗者(**defiers**)(D)。遵从者因为其是合格的而接收处理,总是接受者是指不管其是否合格而接收处理;令不接收者是指不管其合格性如何拒绝处理,违抗者是指其合格却拒绝处理或若不合格却接收处理。倘若不存在违抗者,则只存在三种类型。

处理效应的沃尔德估计量(**Wald estimator**)可由:

$$\mathrm{TE}_{\mathrm{WALD}}=\frac{\mathrm{E}[y_i|z_i=1]-\mathrm{E}[y_i|z_i=0]}{\mathrm{E}[D_i|z_i=1]-\mathrm{E}[D_i|z_i=0]} \tag{25.68}$$

定义,其中,分子被表述成对三种类型处理效应的加权平均,其权数等于成为每一类型的概率,也就是说:

$$\begin{aligned}&\Pr[C]\{\mathrm{E}[y_i|z_i=1,C]-\mathrm{E}[y_i|z_i=0,C]\}\\&\quad+\Pr[A]\{\mathrm{E}[y_i|z_i=1,A]-\mathrm{E}[y_i|z_i=0,A]\}\\&\quad+\Pr[N]\{\mathrm{E}[y_i|z_i=1,N]-\mathrm{E}[y_i|z_i=0,N]\}\\&=\Pr[C]\{\mathrm{E}[y_i|z_i=1,C]-\mathrm{E}[y_i|z_i=0,C]\}\end{aligned}$$

最后一行结果之所以成立，是因为对应于总是接收者与永不接收者的项恒等于 0。式(25.68)的分母遵从的概率为 Pr[C]。因此，有：

$$\mathrm{TE}_{\mathrm{WALD}}=\mathrm{E}[y_{1,i}\mid z_i=1,C]-\mathrm{E}[y_{0,i}\mid z_i=0,C] \tag{25.69}$$

如果将 $\mathrm{TE}_{\mathrm{WALD}}$ 与 LATE 测量进行比较，可以发现，LATE 是关于那些处于参与边缘的子组的处理效应的测量，表示成遵从者。

在实证经济应用中，边际影响是由连续变量变化引起的，这由偏导数来测算，从而很好地得出边际影响，而当原因变量变化是离散的时候，其测算由离散类型代替。因而，以 $\mathbf{x}$ 为条件的边际处理效应(**marginal treatment effect，**记为 **MTE**) 测量被定义成：

$$\mathrm{MTE}=\left.\frac{\partial \mathrm{E}[y\mid \mathbf{x},z]}{\partial \mathrm{Pr}[D=1\mid \mathbf{x},Z]}\right|_{Z=z} \tag{25.70}$$

赫克曼和维特拉西尔(Heckman and Vytlacil，2002)已经证明，ATE、ATET 以及 LATE 都是 MTE 在 Z 支集的不同子集或子总体上取值的平均值。ATE 是 MTE 在 z 的所有支集上的期望值，包括参与率为 0 或 1 的情形。ATET 排除掉没有出现参与的 z 的支集。LATE 是 MTE 在参与率不同的 z 区间上的平均。

25.7.3 *含有异质性处理效应模型中的IV估计*

现在，我们考察允许基于不可观测成分选择与异质性处理效应的模型。这一内容是含有内生处出变量的线性模型，内生处理变量的系数是随机的；参见比约克隆和莫菲特(Björklund and Moffitt，1987)。这类模型是由处理效应经历已处理中不是常值而激发的，伍德里奇(Wooldridge，1997)与赫克曼和维特拉西尔(Heckman and Vytlacil，1998)曾对此研究过。

我们将该模型写成含有结果变量 y_1 的联立方程模型，其中，y_1 依赖于处理变量 y_2。为了简单起见，处理变量 y_2 采用连续的。给定工具 z 与外生变量 x_i，该模型如下：

$$\begin{aligned}y_{1,i}&=(\alpha+v_i)y_{2i}+\mathbf{x}_i'\boldsymbol{\beta}_1+\varepsilon_i \qquad (25.71)\\&=\alpha y_{2i}+\mathbf{x}_i'\boldsymbol{\beta}_1+\varepsilon_i+v_iy_{2i}\\&=v_i\bar{y}_2+\alpha y_{2i}+\mathbf{x}_i'\beta_1+w_i\end{aligned}$$

$$y_{2i}=\gamma z_i+\mathbf{x}_i'\boldsymbol{\beta}_2+\eta_i \tag{25.72}$$

其中，$w_i=\varepsilon_i+v_i(y_{2i}-\bar{y}_2)$。$y_1$ 关于 y_2 变化的边际响应是$(\alpha+v_i)$，对不同个体来说，该值是变动的，因而允许出现异质性处理效应(**heterogeneous treatment effect**)。

假定 $\mathrm{E}[\varepsilon_i\mid \mathbf{x}_i,y_{2i}]=\mathrm{E}[v_i\mid \mathbf{x}_i,y_{2i}]=0$。于是，$\mathrm{E}[\varepsilon_i+v_iy_{2i}\mid \mathbf{x}_i,y_{2i}]=0$，而 $\mathrm{V}[\varepsilon_i+v_iy_{2i}\mid \mathbf{x}_i,y_{2i}]$依赖于 $\mathbf{x}_i$，从而是异质性的。另外，$(\alpha,\ \boldsymbol{\beta}_1)$的最小二乘法估计量是一致的，但不是有效的。这可由 y_2 的假定外生性得到。

其次，考察处理变量是内生的情况。做出下述假设：

$$\mathrm{E}[\varepsilon_i\mid \mathbf{x}_i,z_i]=\mathrm{E}[\eta_i\mid \mathbf{x}_i,z_i]=\mathrm{E}[v_i\mid \mathbf{x}_i,z_i]=0 \tag{25.73}$$

$$\mathrm{E}[\varepsilon_i^2\mid \mathbf{x}_i,z_i]=\sigma_\varepsilon^2;\ \mathrm{E}[v_i^2\mid \mathbf{x}_i,z_i]=\sigma_v^2;\ \mathrm{E}[\eta_i^2\mid \mathbf{x}_i,z_i]=\sigma_\eta^2 \tag{25.74}$$

内生性是通过允许 v 与 η 之间相关而引起的。特别地，假定 $E[v_i|\eta_i]=\rho\eta_i$，如果 (v,η) 是二元正态分布，那么该假定成立。在这些假设下，z 是有效工具，而 $\mathbf{x}$ 是外生的。从 y_1 方程中排除 z 是识别约束。因此，带有工具 $(z,\mathbf{x})$ 的式(25.71)工具变量方程是一个自然估计量。然而，注意到，一致估计的条件是 $E[w_i|\mathbf{x}_i,z_i]=0$。由假设知，$w_i$ 的第一个分量 ε_i 与 z_i 是不相关的；乍看起来，w_i 的第二个分量 $v_i(y_{2i}-\bar{y}_2)$ 好像与 z_i 是相关的。要是这样，Ⅳ估计量就是一致的。不过，可以证明，在上述Ⅳ假设下，估计量是一致的。在论证时，其关键步骤涉及证明 $E[v_iy_{2i}|z_i]=E[v_iy_{2i}]$，伍德里奇(Wooldridge, 1997)通过运用期望迭代定律，已经得出一个结果；因而：

$$\begin{aligned}E[vy_2|z]&=E[E[vy_2|z,\eta]|z] \\ &=E[y_2E[v|z,\eta]|z]=E[\rho\eta y_2|z] \\ &=\rho E[\eta^2|z]=\rho\sigma_\eta^2=E[v\,y_2]\end{aligned} \tag{25.75}$$

给定这里的一些假设，虽然Ⅳ估计量是一致的，但它因为异方差误差而不是有效的。因此，应该使用异方差一致的标准误差。最后，当处理响应是异方差的，我们没有解决估计处理效应对工具选择的敏感性问题。

25.7.4 *非线性模型的内生处理*

当职业培训项目的结果是就业而不是工资，或者是工作安排持续期限时，考察25.3节与25.7节中的分析会如何变化。一种可供选择的方式是，假定处理后的一个显著比例仍是失业的且工资为0，因此，样本是一个具有0与正收入的那些人的混合形式，因而是非正态的。为了处理非线性与非正态性，人们应该怎样推广前面方法呢？

对含有选择的处理与结果的非线性、非正态模型进行估计，是微观经济计量学中常常出现的一个问题。如同线性模型一样，这类模型的主要焦点在于内生处理变量对经济的效应。模型设定由带有因果结构解释的结果方程与对处理变量生成过程建模的其他一些方程组成。就此问题而言，存在两种主要方法：一种是参数方法，依赖于基于似然的(包括贝叶斯)方法，另一种是半参数方法，依赖于GMM或者线性化的Ⅳ方法。

典型设置背景可通过选取的下述一些例子来阐明。在劳动经济学里，宾利和沃克(Bingley and Walker, 2001)考察了丈夫失业持续期限对妻子的离散劳动力供给选择的效应。此处，处理变量是非负的且可能删失的或截取的。皮特和罗森茨韦格(Pitt and Rosenzweig, 1990)研究了内生的幼儿健康状况对他们母亲的每日主要活动的效应；这里处理变量是离散的，而结果是连续的。卡拉斯科(Carrasco, 2001)考察了分娩对妇女劳动力参与的效应。在对照结果模型与生育力有关的情形下，詹森(Jensen, 1999)考察了避孕药物，即离散变量，对分娩之间持续期限，也就是受限因变量的效应。奥尔森和法卡斯(Olsen and Farkas, 1989)研究了生孩子对失学风险的效应。在卫生经济学方面，肯克尔和泰尔扎(Kenkel and Terza, 2001)研究了医生建议(离散的)对酒消费(连续的且非负的)效应。高里森卡拉和汤(Gowrisankaran and Town, 1999)研究了选择医院对死于医院风险的效应。在

卫生经济学中，选择健康保险对健康保健使用(**health care utilization**)的效应，有时被测量成消费变量，有时被看成是某种特定服务类型单位的数量计数，例如医生出诊或住院，这常常利用两部分模型框架加以研究[德布和特里维迪(Deb and Trivedi, 1997)]以及范·奥费姆(van Ophem, 2000)对家庭车辆所有权相对旅游次数效应进行了建模。许多其他例子也可以引用。

这些模型共同具有许多统计特性。第一，处理过程与结果过程都是非正态的及非线性的：多项式的、计数的、离散的或者删失的。第二，在每一种模型中，处理是内生的。最后，研究者经常拥有好的优先选择既有处理的又有结果的特殊参数边际模型的理由。然而，从给定边缘分布到关于处理与结果的联合模型的转变是一个基本步骤，当涉及非正态多变量分布时，该步骤潜在地存在问题。边缘模型常常没有(或非常有约束性)易于处理的多变量对应形式(例如，计数与持续期限模型)。在其他一些模型中，处理与结果源自不同统计族(比如，处理是多项式的，而结果是风险率)，所以不存在解析形式上易处理的多变量分布。由于这一领域中的应用有其特殊性质，所以对此专题不再做任何进一步研究。

25.8 例子：培训对工资的效应

国家支持工作(**National Supported Work, NSW**)示范项目在 20 世纪 70 年代实施，通过随机化实验来测算培训对工资的效应，随机实验意指对某些个体指派接收培训(处理组)，而对另外一些指派不接收处理(对照组)。于是，培训效应能通过直接比较处理后的处理组与对照组的样本均值来进行测算。

如同第 3 章曾讨论的，在社会科学中随机化实验相对稀少。大多数观测样本是使用某些观测到的接收处理的一些个体，而另一些个体则没有接收处理。把已处理组与未处理组进行对比，必须控制观测到的特征以及可能没有观测到特征方面的差异。

为了确定关于观测数据的标准微观经济计量方法的适宜性，拉隆德(Lalonde, 1986)把 NSW 已处理组的结果与那些源自两次国家调查(普查)的对照组进行对比。他获得的结果与把 NSW 已处理组与对照组进行对比的实验结果相差甚远，从而他得出结论，观测方法是靠不住的。

德赫贾和沃赫拜(Dehejia and Wahba, 1999, 2000)利用可供选择的匹配方法重新分析了拉隆德数据子集，他们运用观测数据进行论证，推导结论，这里的观测数据相当接近于来自实验数据的那些情况。在本节，我们利用德赫贾和沃赫拜(Dehejia and Wahba, 1999)的数据来阐明在 25.2～25.5 节介绍的仅仅控制对可观测成分进行选择的一些方法的应用。

25.8.1 *德赫贾与沃赫拜数据*

已处理样本是在 1976～1977 年间接收培训的 185 名男子之一。对照组从 PSID 尚未退休的 55 岁以下的 2 490 名家庭户主中抽取。德赫贾与沃赫拜(Dehejia and Wahba, 1999)称这两个样本为(已处理的)RE74 子样本以及(未处理的)

PSID-1 样本。处理指示变量 D 被定义成 $D=1$，如果接收培训（因而，观测值位于已处理样本中）；以及 $D=0$，如果没有接收培训（从而，观测值位于对照样本中）。

重要变量的概括统计量已由表 25.3 给出。已处理组相当不同于对照组，是与小于高中程序（71%）以及在处理前 1975 年失业（71%）不成比例的黑人（84%）。对培训效应的估计应控制这些差异。

表 25.3　培训影响：已处理组与对照组样本的样本均值[a]

变量	定　义	已处理	控制
AGE	年龄	25.82	34.85
EDUC	受教育年数	10.35	12.12
NODEGREE	当 EDUC<12 时，则为 1	0.71	0.31
BLACK	当民族是黑人，则为 1	0.84	0.25
HISP	当民族是西班牙人，则为 1	0.06	0.03
MARR	当是已婚，则为 1	0.19	0.87
U74	当在 1974 年失业，则为 1	0.60	0.10
U75	当在 1975 年失业，则为 1	0.71	0.09
RE74	1974 年实际工资（1982 年美元）	2 096	19 429
RE75	1975 年实际工资	1 532	19 063
RE78	1978 年实际工资	6 349	21 554
D	若接受培训（处理），则为 1	1.00	0.00
样本量		185	2 490

[a] 数据与德赫贾和沃赫拜（Dehejia and wahba，1999）的表 1 相同。已处理组是 NSW 子样本的 RE74。对照组是未退休的年龄在 55 岁以下男性家庭户 PSID-1 主样本。处理发生在 1976～1977 年。

25.8.2 *控制函数方法*

对培训对工资效应的各种估计已由表 25.4 给出。

关注的结果是处理后工资 RE78。一种可行的培训效应测量是，在 NSW 已处理与 PSID 控制个体之间在 RE78 上的平均差异，得出估计值为 6 349 美元－21 554 美元＝－15 205 美元。这称为处理—对照比较（**treatment-control comparison**）估计量，因为它模仿了实验背景下的分析。

处理—对照比较估计量可等价地被计算，因为它模仿了实验背景下所进行的分析。它等价于计算 RE78 对截距与 D 进行 OLS 回归中的处理标示，当利用组合的处理对照样本变量 D 的系数。

很大的处理估计值会使人误入歧途，因为它大部分反映出两个样本中个体模型方面的差异。即对照样本个体没有控制好。这种差异能够通过引入处理前特征作为回归元而得到控制，并且通过 OLS 估计：

$$\mathrm{RE78}_i=\mathbf{x}_i'\boldsymbol{\beta}+\alpha D_i+u_i,\quad i=1,\cdots,2\,675 \tag{25.76}$$

一旦遵循德赫贾和沃赫拜的线索，把回归元 $\mathbf{x}$ 设定成截距、AGE、AGESQ、EDUC、NODEGREE、BLACK、HISP、RE74 以及 RE75 时，这就导致了更小的估计处理效应 $\hat{\alpha}=218$ 美元。此方法在 25.3.3 节称为控制函数估计量（**control function estimator**）。

25.8.3 差异中差分

第二种方法是前后比较（**before-after comparison**），考虑处理后工资 RE78 与处理前 RE75 之间的差异。若利用已处理组的平均工资，则得到差异估计值 6 349 美元－1 532 美元＝ 4 817 美元。

这个估计值可能导致错误结论，因为它反映出该时期的所有变化，诸如经济改进而不只是培训。在 25.5 节曾考虑的差异中差分估计量（**difference-in-differences estimator**）另外计算对照组中类似的量，即 21 554 美元－19 063 美元＝2 491 美元，并利用这作为工资期间相关未处理变化的测量，所以仅仅因为处理而随时间变化的是 4 817 美元－2 491 美元＝2 326 美元。

可以证明，DID 估计量等价于对 OLS 回归：

$$\mathrm{RE}_{it}=\phi+\delta \mathrm{D78}_{it}+\gamma\alpha \mathrm{D}_i+\alpha \mathrm{D78}_{it}\times \mathrm{D}_i+u_i,\quad i=1,\cdots,2\ 675,\ t=75,78 \tag{25.77}$$

此处，$\mathrm{RE}_{i,75}$ 表示处理前时期的工资，而 $\mathrm{RE}_{i,78}$ 表示处理后时期的工资，因此，该回归是带有 5 350 个工资观测值的回归。指示变量 $\mathrm{D78}_{it}$ 在处理后时期为 1，如果个体位于已处理样本之中，那么指示变量 D_i 等于 1，而交叉项 $\mathrm{D78}_{it}\times \mathrm{D}_i$ 对于处理后时期的已处理个体来说等于 1。

更一般地，式(25.77)中截距能用 $\mathbf{x}_{it}'\boldsymbol{\beta}$ 来代替。在此例子中并不会产生差异，因为回归元是时常值的，所以 $\mathbf{x}_{it}=\mathbf{x}_i$。

这一方法能用于重复横截面数据（参见 22.6.2 节），因为它不需要已处理组与对照组中的个体在 1975 年和 1978 年中都是可观测到的。

25.8.4 简单倾向得分估计

第三种方法是把已处理个体的结果 RE78 与 RE78 的反事实预测相对比，如果相同的已处理个体实际上没有接收处理。初始 15 205 美元的处理—对照估计是一种过分简化例子，它用作对照组（21 554 美元）中 RE78 的反事实平均值。更好的反事实能通过设定回归模型来生成。例如，如果已处理，回归(25.76)就设定 $\mathrm{E}[\mathrm{RE78}|\mathbf{x}]$等于 $\mathbf{x}'\boldsymbol{\beta}+\boldsymbol{\alpha}$，如果未处理，设定反事实 $\mathbf{x}'\boldsymbol{\beta}$。这既对回归元 $\mathbf{x}$ 的效应施加了约束，又对处理效应施加了约束，约束是以 $\mathbf{x}$ 为条件的，假定对于不同个体而言是常值。

处理效应文献强调并不依赖于如此强假设的反事实。一种明显方法是，将已处理及未处理个体与 $\mathbf{x}$ 的相同值加以比较，但如果几个回归元被认为是有意义的，而且这些回归元取一系列不同的值，那么对回归元匹配（**matching on regressors**）是不可解的。

然而，给定 25.3 节与 25.4 节详述的假设，对倾向得分匹配（**match on the propensity score**）就足够了，这里倾向得分匹配被定义成处理的条件概率 $\Pr[D=1|\mathbf{x}]$。对此例来说，我们只利用初始 1975 年的数据可估计 logit 模型：

$$\Pr[D_i=1|\mathbf{x}_i]=\Lambda(\mathbf{x}_i'\boldsymbol{\beta}),\quad i=1,\cdots,2\ 675 \tag{25.78}$$

其中，由 14.2 节知，$\Lambda(z)=e^z/(1+e^z)$，遵循德赫贾和沃赫拜（Dehejia and Wahba，1999）线索，选择回归元为 AGE、AGESQ、EDUC、EDUCSO、NODEGREE、BLACK、HISP、MARR、RE74、RE75、RE74SQ、RE75SQ 以及 U74*BLACK。

图 25.3 画出，处理后工资 RE78 与倾向得分，分别各自绘制出已处理样本及对照组样本。当仅仅考虑倾向得分（x 轴）时，很明显，对照组中绝大多数部分观测值具有非常小的倾向得分，期望结果已由 25.3 给出数据，即已处理个体是成比例的黑人、失业、受教育年数少的个体。

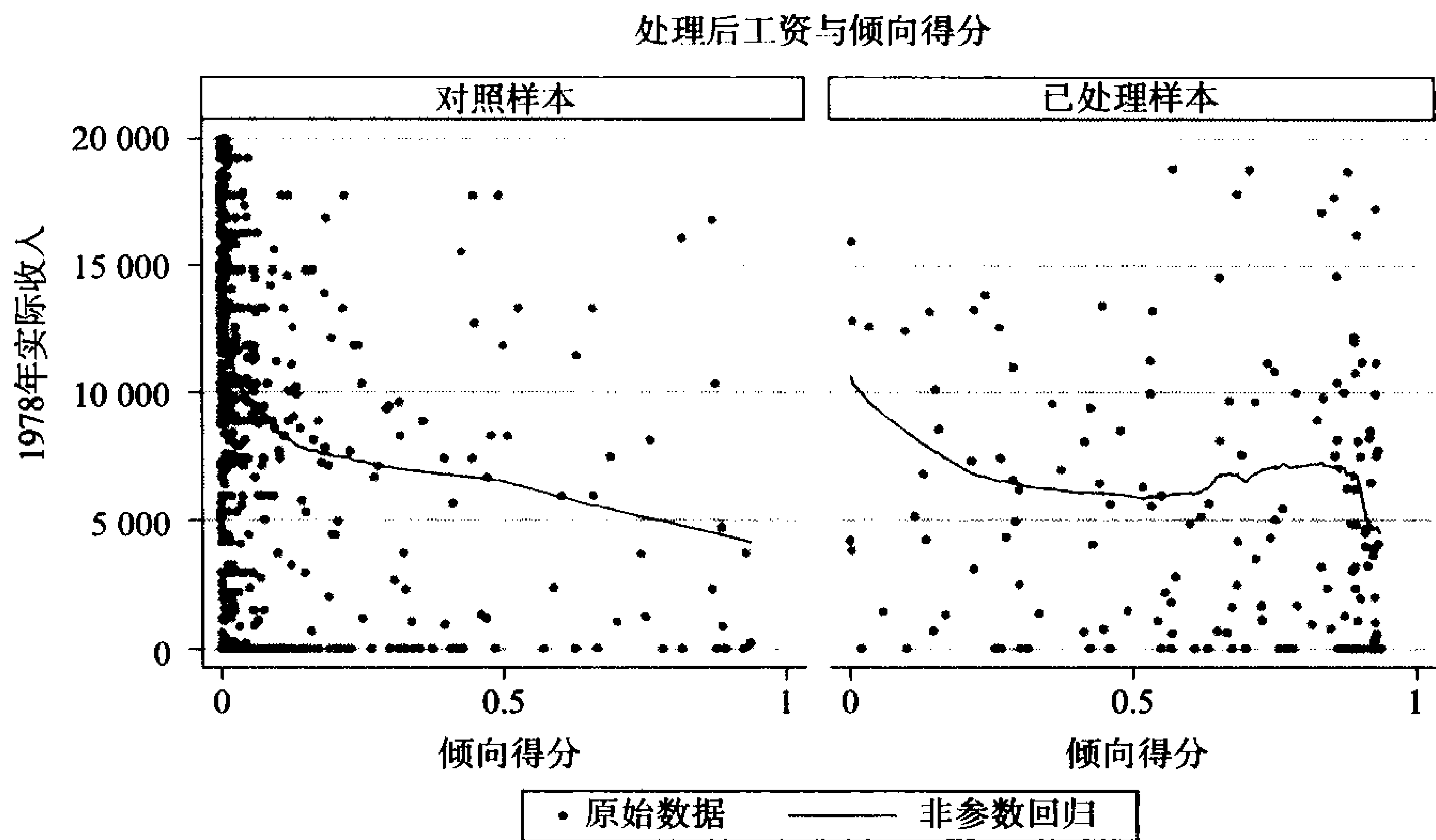


图 25.3 培训影响。依据处理状态，处理后的收入与倾向得分的散点图。这里只包含有相同倾向得分的那些观测值。为了方便观察，收入超过 20 000 美元的观测值被排除在散点图之外，尽管非参数回归包括这些观测值。

转到处理后结果 RE78（y 轴），可以发现，处理效应被估计成给定已处理个体（$D=1$）与含有相同（预测）倾向得分对照样本个体之差。图 25.3 中每组都包括了 RE78 对倾向得分的拟合非参数回归。在相当大的倾向得分范围内，处理效应小于 1 000 美元，尽管对于倾向得分 0.80 左右来说，它相当大且为正的。

存在许多实施这种将个体与类似倾向得分进行比较的方法，然后对所有已处理个体进行平均。一种策略是，将已处理个体与对照样本中具有最接近倾向得分的个体加以匹配。此方法在 25.4.4 节称为最近邻匹配。比较简单的策略是，通过倾向得分记为 $p(\mathbf{x})$，将数据分成层，并设反事实条件陈述是对照组 ER78 的组内平均值。例如，如果已处理观测值具有倾向得分 $p(\mathbf{x})=0.35$，那么反事实条件陈述就是关于对照组的 RE78，观测值满足 $0.30\leqslant p(\mathbf{x})\leqslant 0.40$ 的 $p(\mathbf{x})$ 平均值。于是，总效应是 $\sum_s w_s(\overline{\mathrm{RE78}}_{s,D=1}-\overline{\mathrm{RE78}}_{s,D=0})$，其中，$\overline{\mathrm{RE78}}_{s,D=1}$ 与 $\overline{\mathrm{RE78}}_{s,D=0}$ 分别表示已处理观测值与未处理观测值，而其权数 w_s 等于每一层中已处理观测值的比例。一种简单分层方案使用，比如说 10 等份空间分层，满足 $0.0<p(\mathbf{x})\leqslant 0.1$，$0.1<p(\mathbf{x})\leqslant 0.2$ 等。这在 25.4.4 节称为分层匹配。这种方法应该被限制在已处理样本与对照样本交叠的倾向得分情形上，参见 25.4.3 节。此处，已处理样本的倾向得分范围从 0.000 5 到 0.942 0，而对照样本的倾向得分范围从 0.000 0 到0.937 1，

导致了 1 423 个对照组个体与 8 个已处理个体被省略掉。所得到的估计总效应是 995 美元，已在表 25.4 中给出。

表 25.4　培训影响：处理效应的各种不同估计值

方　法	定　　义	估计值	标准误差[a]
处理—对照比较	$\overline{\text{RE78}}_{D=1}-\overline{\text{RE78}}_{D=0}$	−15 205	656
控制函数估计量	源自 OLS 回归(25.76)$\hat{\alpha}$	218	768
前后比较	$\overline{\text{RE78}}_{D=1}-\overline{\text{RE75}}_{D=1}$	4 817	625
差异中差分	源自 OLS 回归(25.77)$\hat{\alpha}$	2 326	749
倾向得分	参见 25.8.4 节	995	—

[a] 前四个估计值的标准误差是利用来自适当 OLS 回归的异方差一致标准误差计算出来的。

25.8.5　*利用倾向得分匹配*

如同 25.4 节提及的，其他一些匹配策略包括半径匹配与核匹配，它们同样相对容易实施。本章余下内容详细讨论这些方法及其他方法，特别强调倾向得分方法。

拟合倾向得分

拟合倾向得分可利用分别源自德赫贾和沃赫拜(Dehejia and Wahba, 1999)以及德赫贾和沃赫拜(Dehejia and Wahba, 2002)的两个不同 logit 设定来获得。关于倾向得分的设定已在表 25.6 底部详细给出。仅在违离德赫贾和沃赫拜(Dehejia and Wahba, 1999, 2002)情况下，我们的 logit 模型包括了常值项。为了节省篇幅，不阐述系数估计，但表示预计的符号模式。

匹配算法与平衡

一个重要的实际应用问题是，选择一种基于倾向得分的适当匹配算法，使得平衡条件(25.9)满足。德赫贾和沃赫拜(Dehejia and Wahba, 2002，第 16 页)提供了以简洁 logit 模型对 $p(\mathbf{x})$进行估计的算法。该算法原理如下。依照 $\hat{p}(\mathbf{x})$对数据进行分类。对样本观测值加以分层，使得层内关于已处理单元与控制单元的 $\hat{p}(\mathbf{x})$都很接近。例如，起初使用等范围的粗格子。对于每一层来说，已处理单元与未处理单元之间均值相等就每个协变量而言都要加以检验。如果不存在统计上的显著差异，那么回归元在已处理组与未处理之间是平衡的，从而人们可以停止。对于某些层来说，如果不存在平衡，那么就非平衡层(**unbalanced stratum**)而言，使用更精细格子来达到平衡。若存在许多非平衡层，则运用包含回归元之间交叉项及较高阶项的改进设定来重新估计最初的 logit 模型。

利用贝克尔和市野(Becker and Ichino, 2002)的软件，德赫贾和沃赫拜(Dehejia and Wahba, 2002)算法可用于计算倾向得分。在所有注意到的情况下，倾向得分计算被限制在对平衡性质(**balancing property**)进行检验的共同支撑区域，该检验利用那些倾向得分位于已处理单元与对照组单元的倾向得分支集的交集中的观测值。这种限制显著地减少了最初样本。就德赫贾和沃赫拜(Dehejia and Wahba, 2002)设定而言，对照组容量从 2 490 单元减少到 1 086 单元。

表 25.5 给出了在实施平衡之后，通过刚才概述方法完成的不同组中一系列已处理单元与控制单元。报告结果与德赫贾和沃赫拜(Dehejia and Wahba, 2002)的

那些不同,因为后者从不以共同支撑区域为基础的 NSW - PSID 合成样本中排除掉对照单元,只是以样本单元估计倾向得分是否小于已处理单元的估计倾向得分最小值为基础。该表显示,若与其他组相比较,已处理单元与对照单元的比例,就第一组而言是相当小的。

表 25.5 培训影响:利用德赫贾和沃赫拜(1999)设定的关于已处理单元与对照单元的倾向得分分布[a]

最小值 $\hat{p}(\mathbf{x})$	已处理的	未处理的	总数
0.000 364	9	960	969
0.10	10	56	66
0.20	14	33	47
0.40	24	22	46
0.60	33	7	40
0.80	95	8	103
总计	185	1 086	1 271

[a] 例如,从第二行知,倾向得分位于 10 个已处理个体与 56 个未处理个体的 0.10 与 0.20 之间。

对德赫贾和沃赫拜(Dehejia and Wahba, 1999)的设定可做出类似检验,简单起见,这里就不列表显示了,但仍会产生类似结果。对照组具有 1 146 个观测值。于是,对于分组 $\hat{p}(\mathbf{x})$来说,有界值是 0.000 652 6, 0.05, 0.10, 0.20, 0.40, 0.60 以及 0.80。

通过匹配方法对 ATET 估计

关于各种匹配方法结果的选择已概括归纳在表 25.6 中。就德赫贾和沃赫拜(Dehejia and Wahba, 2002)设定而言,ATET 的最近邻估计值是 2 385 美元,而就

表 25.6 培训影响:ATET 的估计值

匹配方法	已处理数	对照组数	ATET	标准差	占 1 794 美元%[e]
德赫贾和沃赫拜(2002)设定[a]					
最近邻	185	53	2 385	1 209[c]	133
半径, $r=0.001$	54	517	−7 815	1 118[d]	−436
半径, $r=0.000\,1$	24	92	−9 333	2 282[d]	−520
半径, $r=0.000\,01$	15	19	−2 200	2 986[d]	−120
分层	185	1 086	1 452	1 041[c]	81
核	185	1 058	1 309	975[c]	73
德赫贾和沃赫拜(1999)设定[b]					
最近邻	185	57	560	1 098[c]	31
半径, $r=0.001$	57	583	−9 358	997[d]	−522
半径, $r=0.000\,1$	27	76	−7 847	2 066[d]	−437
半径, $r=0.000\,01$	16	13	223	4 551[d]	12
分层	185	1 146	2 156	814[c]	120
核	185	1 146	1 518	890[c]	85

[a] logit 模型:Pr[treat=1]=h(CONSTANT, AGE, AGE², EDU, EDU², MARRIED, NODEGREE, BLACK, HISPANIC, RE74, RE74², RE75, U74, U75, U74*HISPANIC)。

[b] logit 模型:Pr [treat =1]=h(CONSTANT, AGE, AGE², EDU, EDU², MARRIED, NODEGREE, BLACK, HISPANIC, RE74, RE74², RE75, RE75², RE74*RE75, U74*BLACK)。

[c] 带有 200 个复制的自助法标准误差。

[d] 解析标准误差。

[e] ATET/1794×100。

德赫贾和沃赫拜(Dehejia and Wahba, 1999)设定而言,它的 ATET 最近邻估计值大致是 560 美元。分层与核匹配的效果同样可以是混杂的,ATET 的估计值从 1 452 美元到 2 156 美元。

为了比较起见,德赫贾和沃赫拜(Dehejia and Wahba, 2002)ATET 估计值已由表 25.7 重新给出。我们还注意到,处理效应的基准估计是 1 794 美元。它可通过 RE78 对既有参与者又有非参与者的 NSW 样本的德赫贾和沃赫拜(Dehejia and Wahba, 2002)形式进行回归而求出。很明显,此表中报告的 ATET 估计值显著地不同于德赫贾和沃赫拜(Dehejia and Wahba,2002)的那些结果,并且不同于基准实际实验估计值。对德赫贾和沃赫拜(Dehejia and Wahba, 2002)设定来说,其最近邻估计量非常接近于基准估计,而依照缩减偏倚来说,甚至比德赫贾和沃赫拜(Dehejia and Wahba, 2002)的结果更好。

表 25.7　培训评估:德赫贾和沃赫拜(2002)的 ATET 估计

匹配方法	ATET	标准误差
最近邻	1 890	1 202
半径,$r=0.001$	1 824	1 187
半径,$r=0.000\,1$	1 973	1 191
半径,$r=0.000\,05$	1 928	1 196
半径,$r=0.000\,01$	1 893	1 198

对于分层估计与核估计来说,偏倚是较大的。对于半径匹配估计量来说,这种偏倚更粗糙,并且给出处理效应的负估计值,与德赫贾和沃赫拜(Dehejia and Wahba, 2002)利用测径匹配求出的正估计值相反。我们的半径匹配与德赫贾和沃赫拜(Dehejia and Wahba, 2002)的测径匹配之间的差异在于后者方案,当给定已处理单元没有匹配到给定测径之内时,匹配便与给定测径之外最近的比较单位进行匹配。在这种情况下,就忽略没有匹配到预先设定半径上的已处理单元。这阐明了匹配估计量对假设的敏感性。

就各种不同设定而言,给定表 25.6 的最后一列,ATET 估计值的稳健性可依据 ATET 与基准估计之比来计算。除分层匹配估计量以外,其余比率变化对两种设定来说变化更大。例如,在德赫贾和沃赫拜(Dehejia and Wahba, 2002)设定中,最近邻估计量是基准估计量的 133%,但在德赫贾和沃赫拜(Dehejia and Wahba, 1999)设定中,最近邻估计量却仅仅是基准估计量的 31%。类似地,除核估计量之外,ATET 估计对所用倾向得分都是敏感的。

匹配方法是否起作用,依赖于有关已处理组与未处理组的倾向得分模型的适合性[德赫贾和沃赫拜(Dehejia and Wahba, 2002)]。不过,很明显,在方法与倾向得分模型之间存在着交互作用。

25.9　文献注释

匹配方法与差异中差分方法对项目评估的早期经济应用,包括阿申费尔特(Ashenfelter,1978)以及阿申费尔特和卡德(Ashenfelter and Card, 1985)。处理

评估是当今经济计量学研究中一个相当活跃且迅猛发展的领域。

25.2 安格里斯特等人(Angrist et al., 1996)在医学与经济计量学文献方面的概念及术语之间,给出了一种有益的联系。

25.3 赫克曼和罗布(Heckman and Robb, 1985)考察了存在选择条件下,各种数据背景下项目影响的估计。还可参见比约克隆和莫菲特(Björklund and Moffitt, 1987)。赫克曼和霍茨(Heckman and Hotz, 1989)同样非常有说服力地讨论了人们需要在几种设定检验结果条件下评估它们的稳健性,并计算选择偏倚的影响。例如,他们建议使用多重比较组来计算建立在单个对照组基础上结果的敏感性。这方面的早期工作大部分是参数方法。最近,大多数非参数方法也得到了运用。

25.4 赫克曼、市村和托德(Heckman, Ichimura and Todd, 1997)以及赫克曼等人(Heckman et al., 1998)研究并应用了匹配估计量。涉及以倾向得分为条件的重要结果是由罗森鲍姆和鲁宾(Rosenbaum and Rubin, 1983,定理 2)给出的。利用估计倾向得分对 ATE 进行有效估计是由平野、英伯斯和里德(Hirano Imbens and Ridder, 2003)给出的。德赫贾和沃赫拜(Dehejia and Wahba, 2002)将倾向得分方法应用到拉隆德(Lalonde, 1996)数据集变形上。实验数据可与源自 CPS 与 PSID 的观测值相匹配。史密斯和托德(Smith and Todd, 2004)利用倾向得分重新分析了由德赫贾和沃赫拜使用的数据。他们强调与可供选择倾向得分估计量有关的偏倚,突出在偏倚最小化时高质量数据的重要性。贝克尔和一野(Becker and Ichino, 2002)曾提供某些倾向得分匹配估计量的综述。他们还给出 STATA 编程集合加以阐述说明,用于对 ATET 进行估计。《经济学季刊》(*Quarterly Journal of Economics*)2004 年 2 月刊包含了匹配经济计量学研讨会专集。

25.6 哈恩、托德和范德克劳(Hahn, Todd and Van der Klaauw, 2001)在弱假设下,分析了 RD 模型中处理效应的识别问题。

25.7 英伯斯和安格里斯特(Imbens and Angrist, 1994)分析了 LATE 估计量的性质。安格里斯特等人(Angrist et al., 1996)对Ⅳ方法的运用加以讨论,给出与处理影响的 LATE 测量的联系。该论文还给出对Ⅳ估计量的各种层次观点以及文献联系的重要讨论,也可参见赫克曼(Heckman, 1997)。安格里斯特(Angrist, 2001)对在含有非正态结果的非线性结果模型中处理内生虚拟变量的某些简单策略加以讨论。此论文还对线性化Ⅳ方法的优缺点给出一个评述。在一些竞争方法之间,对于最有前途方法的看法难以达成共识。赫克曼、托拜厄斯和维特拉西尔(Heckman, Tobias and Vytlacil, 2003)发展了潜变量框架内的处理效应估计量。维拉和费尔贝克(Vella and Verbeek, 1999)将Ⅳ方法与控制函数方法加以对比,包括选择偏倚校正项。

习 题

25-1 [改编自赫克曼(Heckman, 1996)。]考察处理—对照模型 $y=\mathbf{x}'\beta+\alpha d+\varepsilon$,其中,$d$ 表示二值指示变量,当处理是随机指派时,$d=1$;当处理不是指派(同样不是随机的)时,$d=0$。

(a) 随机化处理是识别 α 的充分条件吗?

(b) 随机化处理是识别 α 与 $\boldsymbol{\beta}$ 的充分条件吗?

25-2 在上一个问题中,随机化涉及处理。这里,我们考察关于接收处理的随机适宜性。现在,$e=1$ 意味着对个体做出随机指派,而 $e=0$ 意味着对个体不做出随机指派。证明,在此情况下,给定 $\Pr[d=1|\mathbf{x}]\neq 0$,处理效应可由 $\mathrm{E}[y|e=1,\mathbf{x}]-\mathrm{E}[y|e=0,\mathbf{x}]/\Pr[d=1|\mathbf{x}]$ 给出。

25-3 考察非线性处理结果模型 $\mathrm{E}[y|\mathbf{x},d]=\exp(\mathbf{x}'\boldsymbol{\beta}+\alpha d)$,其中,$d$ 表示二值处理指示变量。假定可以利用($\boldsymbol{\beta}$, α)的一致估计值,并估计协方差矩阵 $\hat{\mathrm{V}}[\hat{\boldsymbol{\beta}},\hat{\alpha}]$。假定估计量是渐近正态的。请概述 ATE 估计参数及其渐近方差的自助法或蒙特卡罗算法,给定($\mathbf{x}_i$, d_i),$i=1,\cdots,N$。

25-4 考察非线性处理结果模型 $\mathrm{E}[\ln y|\mathbf{x},d]=\mathbf{x}'\boldsymbol{\beta}+\alpha d$,其中,$d$ 表示二值处理指示变量。假定可以利用($\boldsymbol{\beta}$, α)的一致估计值,并估计协方差矩阵 $\hat{\mathrm{V}}[\hat{\boldsymbol{\beta}},\hat{\alpha}]$。假定依据 y 而不是 $\ln y$ 对 ATE 进行估计感兴趣。请提出一种估计方法,并讨论它的一致性。

25-5 在本章,经验例子使用 PSID 对照组与 NSW 处理组。德赫贾和沃赫拜(Dehejia and Wahba, 2002)使用了两个对照组。存在另一种建立在 CPS 基础上的可利用对照组。本题要求你利用 CPS 对照组代替 PSID 样本重复报告计算。

(a) 生成类似于表 25.3 的表格。依据年龄、民族、受教育程度以及处理前工资,将 NSW 组与 CPS 对照组进行比较。

(b) 如同 25.8 节所做的,利用估计倾向得分考察处理组与对照组之间的差异。利用 25.8.4 节方法,估计 NSW - CPS 合成样本的倾向得分,一旦以线性方式并入协方差,同时带有较高阶项时,像德赫贾和沃赫拜(Dehejia and Wahba, 2002)那样。假如忽略那些倾向得分小于处理单元最小值的比较单元,利用直方图对两个倾向得分的集合加以对比。对位于不同倾向得分区间("箱子")中含有比较单元的匹配程度给出评述。

(c) 利用 25.8.4 节与 25.8.5 节曾经阐述并实施的匹配方法(尤其是最近邻、分层或区间匹配、核匹配以及半径匹配),建立一个类似于表 25.6 的表格。对 ATET 估计值加以评述,并将它们与那些建立在 PSID 基础上的比较组加以对比。

26 测量误差模型

26.1 引 论

在经济计量学领域，测量误差问题随处可见。就微观经济计量学而言，测量误差问题的一个共同来源是对调查问题的不正确回答、正确回答的错误记录，以及一个正确测量变量用作另一个理论上有效却观测不到变量的代表（比如，用观测收入代表"正常收入"）。探寻敏感信息问题可能引起部分回答或错误回答。也就是说，当不可观测变量（或潜在变量）被代表变量所代替时，由不可观测变量（或潜在变量）引致的测量误差。

这里举几个例子。考虑研究收入问题时，对性别偏倚存在进行检验。一种明显方法是，一旦控制了各类资格证书、年龄、经历等，将收入测量对性别分类变量回归。可是，最有关的变量可能是个体在职效率，该变量不可能被直接观测到，从而要用其代表变量。因此，测量误差会对性别歧视推断产生影响，这是一个重要问题。研究个体问题时，要考虑对商品及服务的需求，诸如"经济成本"或"全价服务"特征概念。不过，这类概念在出版数据中几乎难以直接测算出，因而必须用经济计量学先验模型估计出来。可是，对此类数据测算必受限于误差。

本书讨论的模型几乎难以避免测量误差问题。二值结果内生或外生变量都潜在地受限于分类误差，源于追溯调查的过渡数据或计数数据均受到回忆误差影响；相对质朴变量，比如小时工资与家庭开支的数据，被故意夸大或报告误差所扭曲。与总量数据不同，汇总可能导致与测量误差的某种相互抵消，但对于个体层面数据来说，测量误差持续存在。

本章第一部分研究测量误差的后果以及用于补救后果的估计策略。这里既讨论线性模型，又讨论非线性模型。尽管更为现实的方式是，承认此类问题经常与其他问题交织在一起出现，但为了解释方便起见，假定经济计量学所面临的问题仅是测量误差。

更宽泛地讲，测量误差的后果是对关注参数识别的失败。解决该问题是极其复杂的。一种方法是，考虑直接省略模型的有关变量，或用代表变量代替真实测量。除一些极端情况之外，都不能这样做，至少存在两个重要原因。首先，若变量处于关注焦点，则省略会产生严重的省略变量偏倚，因此人们是用一种类型问题代

替另一种类型问题，识别仍是不可能的。其次，在线性回归中，倘若测量误差是随机的且与真实回归元独立，则运用潜在变量代表所得到的渐近偏倚比从模型中直接省略潜在变量所产生的偏倚要小[麦卡勒姆(McCallum，1972)]。倘若忽略潜在变量，则会导致不好的估计。不过，运用代表变量仍将得出非一致估计，尽管该偏倚较小。

解决测量误差问题的基本观点是，重新找到潜在变量参数，并识别模型。人们必须拥有关于测量误差的额外假设形式的外来信息或者获得额外数据，而且在做出似乎合理的假设之后，运用这些信息。这是一种十分流行的方法。不过，当没有额外数据可以利用时，就要对经济计量模型提出一种好的可供选择形式。

测量误差会产生潜在的相当严重的后果，这是因为在许多情况下，测量误差使回归参数不可识别。例如，卡德(Card，2001)曾再次考察受教育对收入系数的经验证据，他发现，典型向下偏倚 25%～35%。测量误差的确切后果依赖于模型函数形式、误差是如何进入模型里的(比如是加法形式，还是乘法形式)，以及正在研究的数据结构。解决因测量误差而产生的问题，典型地需要将额外信息引入模型，或者是额外数据形式，或者是额外假设。

本章分别以线性模型与非线性模型各自分开的方式讨论测量误差，无疑这是一种简便的组织安排，然后考察特殊情况。26.2 节与 26.3 节探讨线性回归。26.4 节内容涵盖非线性回归。26.5 节讨论一些蒙特卡罗例子。由线性模型引出基本的直觉观念，为认识非线性模型提供一个有益的基础。在任何情况下，较明确的结果通常都是针对特定模型而得到的。

26.2　线性回归的测量误差

回归元的测量误差也称为变量误差(**error-in-variables**)，尽管测量误差具有零均值，但因为测量误差会使 OLS 估计量出现非一致性，故它是一个重要专题。回归元的测量误差，经常被说成引起偏倚，但我们却使用比较强的术语 —— 非一致性，这是因为当样本量趋于无穷大时，此偏倚没有消失。

测量误差模型的范围非常广泛，涵盖了下述一些情况：测量误差会影响到右边变量(“回归元”)或左边变量(“结果”)，或者对左右两边变量都产生影响。豪斯曼(Hausman，2001)将它们称为“源于右边的问题”与“源于左边的问题”。对于后者，通常称为经典变量误差模型，关注的关系，则是结果 $\mathbf{y}$ 与协变量$(\mathbf{W},\mathbf{X}^*)$之间关系，其中，$\mathbf{W}$ 表示没有误差的测量，而 $\mathbf{X}^*$ 表示不可观测的，但 $\mathbf{X}^*$ 却有代表值可以利用，将其记为 $\mathbf{X}$。关注问题是，$\mathbf{y}$ 与$(\mathbf{W},\mathbf{X})$之间的估计关系是否提供了推断 $\mathbf{X}^*$ 的一个满意基础。

统计文献很容易区分测量误差模型的函数方法与结构方法。若 $\mathbf{X}^*$ 表示真实不可观测的协变量，则函数方法将这些协变量处理成未知固定常值(参数)。而结构方法则将这些协变量处理成随机变量。卡罗尔、鲁珀特和斯特凡斯基(Carroll，Ruppert，and Stefanski，1995)进一步区分了函数建模法与结构建模法，函数建模法意指不管协变量是固定的还是随机的，对于 $\mathbf{X}$ 的情况均能做出唯一最小假设，而

结构建模法则对 X 分布做出参数假设。函数测量误差模型是带有无限多冗余参数模型的例子，因此，极大似然法有众所周知的缺陷（面板数据章节曾经讨论过）。经济计量学文献对这种区分缺乏共识。

在应用时，出现的非一致性的程度是相当大的。在对个体收入决定因素进行经济计量研究时，对测量误差以及控制它们的方法尤其要进行广泛讨论。

26.2.1 *经典测量误差模型*

标准测量误差模型具有连续因变量 y，该因变量 y 是 K 个真实回归元 $\mathbf{x}^*$ 的一个线性函数。若 y 的加法测量误差与回归元不相关，则加法测量误差就不会产生任何问题，这是因为它被吸收到方程误差之中。当 $\mathbf{x}^*$ 是可观测的，则通过 y 对 $\mathbf{x}^*$ 的普通最小二乘法回归

$$y_i=\mathbf{x}_i^{*\prime}\boldsymbol{\beta}+u_i$$

能一致地估计出参数，其中，u_i 是 iid $[0,\sigma^2]$。否则，观测数据是 $\mathbf{x}\neq\mathbf{x}^*$，$y$ 要对 $\mathbf{x}$ 而不是 $\mathbf{x}^*$ 进行回归。假定真实回归元与观测回归元之间的关系是：

$$\mathbf{x}_i=\mathbf{x}_i^*+\mathbf{v}_i,\quad i=1,\cdots,N \tag{26.1}$$

其中，加法测量误差被假定成服从下面分布：

$$\mathbf{v}_i\sim[\mathbf{0},\boldsymbol{\Sigma}_{\mathbf{vv}}] \tag{26.2}$$

不可观测真实回归元被假定具有零均值，原因在于将变量测量成偏离均值形式，且具有方差矩阵：

$$\mathrm{V}[\mathbf{x}_i^*]=\boldsymbol{\Sigma}_{\mathbf{x}^*\mathbf{x}^*} \tag{26.3}$$

注意，$\mathbf{x}$ 是 $\mathbf{x}^*$ 的无偏估计值，因为测量误差被假定成有零均值。若假定测量误差既与 $\mathbf{x}^*$ 独立，又与回归误差 u 独立，则有：

$$\mathrm{E}[\mathbf{v}_i|\mathbf{x}_i^*]=\mathrm{E}[\mathbf{v}_i|u_i]=0 \tag{26.4}$$

26.2.2 *OLS 的非一致性*

考察测量误差的后果，将假定的经典测量误差模型的数据生成过程用矩阵记号写成：

$$\begin{aligned}\mathbf{y}&=\mathbf{X}^*\boldsymbol{\beta}+\mathbf{u}\\ \mathbf{X}&=\mathbf{X}^*+\mathbf{V}\end{aligned} \tag{26.5}$$

是有益的，其中，方程误差 u 遵从条件 $\mathrm{E}[\mathbf{u}|\mathbf{X}^*]=\mathbf{0}$ 与 $\mathrm{E}[\mathbf{uu}'|\mathbf{X}^*]=\sigma^2\mathbf{I}_N$。将第二个方程代入第一个方程，得到：

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+(\mathbf{u}-\mathbf{V}\boldsymbol{\beta}) \tag{26.6}$$

由于误差项 $(\mathbf{u}-\mathbf{V}\boldsymbol{\beta})$ 经由测量误差而与回归元 $\mathbf{X}$ 相关，所以 $\mathbf{y}$ 对 $\mathbf{X}$ 的 OLS 回归会得出 $\boldsymbol{\beta}$ 的非一致估计。

正式地讲，我们有：

$$\begin{aligned}\text{plim}\, N^{-1}\mathbf{X}'(\mathbf{u}-\mathbf{V}\boldsymbol{\beta}) &= \text{plim}\, N^{-1}(\mathbf{X}^* + \mathbf{V})'(\mathbf{u}-\mathbf{V}\boldsymbol{\beta}) \\ &= -\boldsymbol{\Sigma}_{\text{vv}}\boldsymbol{\beta} \\ &\neq \mathbf{0}\end{aligned}$$

这里使用了 $N^{-1}\mathbf{V}'\mathbf{V}=N^{-1}\sum_i \mathbf{v}_i\mathbf{v}_i'$ 以及 $\mathbf{v}_i$ 是 iid $[\mathbf{0},\boldsymbol{\Sigma}_{\text{vv}}]$。这是非一致性的根本来源。现在：

$$\begin{aligned}\text{plim}\, N^{-1}\mathbf{X}'\mathbf{X} &= \text{plim}\, N^{-1}(\mathbf{X}^* + \mathbf{V})'(\mathbf{X}^* + \mathbf{V}) \\ &= \boldsymbol{\Sigma}_{\text{x}^*\text{x}^*} + \boldsymbol{\Sigma}_{\text{vv}}\end{aligned}$$

其中，用到了 $\mathbf{x}_i^*$ 的 iid 性质，$\mathbf{x}_i^*$ 均值为 0 且 $\text{V}[\mathbf{x}_i^*]=\boldsymbol{\Sigma}_{\text{x}^*\text{x}^*}$。此外，有：

$$\begin{aligned}\text{plim}\, N^{-1}\mathbf{X}'\mathbf{y} &= \text{plim}\, N^{-1}(\mathbf{X}^* + \mathbf{V})'(\mathbf{X}^*\boldsymbol{\beta} + \mathbf{u}) \\ &= \boldsymbol{\Sigma}_{\text{x}^*\text{x}^*}\boldsymbol{\beta} \\ &\neq \mathbf{0}\end{aligned}$$

因而，当应用斯卢茨基定理(附录 A，定理 A. 3)，得到：

$$\begin{aligned}\text{plim}\, \hat{\boldsymbol{\beta}} &= (\text{plim}\, N^{-1}\mathbf{X}'\mathbf{X})^{-1}\text{plim}\, N^{-1}\mathbf{X}'\mathbf{y} \\ &= (\boldsymbol{\Sigma}_{\text{xx}})^{-1}(\boldsymbol{\Sigma}_{\text{xx}} - \boldsymbol{\Sigma}_{\text{vv}})\boldsymbol{\beta} \\ &= \boldsymbol{\beta} - (\boldsymbol{\Sigma}_{\text{x}^*\text{x}^*} + \boldsymbol{\Sigma}_{\text{vv}})^{-1}\boldsymbol{\Sigma}_{\text{vv}}\boldsymbol{\beta}\end{aligned} \tag{26.7}$$

很明显，只要存在测量误差且 $\boldsymbol{\Sigma}_{\text{vv}} \neq \mathbf{0}$，就会使 OLS 是非一致的。

为了后面参考方便，注意到，当我们可以利用 $\boldsymbol{\Sigma}_{\text{vv}}$ 的一致估计值，将其记为 $\mathbf{S}_{\text{vv}}$，并且 $(\mathbf{X}'\mathbf{X}-\mathbf{S}_{\text{vv}})$ 是正定的，就能计算调整后的最小二乘估计量 $\hat{\boldsymbol{\beta}}_a=(\mathbf{X}'\mathbf{X}-\mathbf{S}_{\text{vv}})^{-1}\mathbf{X}'\mathbf{y}$。该公式还能用于研究测量误差方差的假设值对最小二乘估计量的影响。

26.2.3 *纯量回归元的测量误差*

教科书通常考虑这种模型的特殊情况，即考察如下情况：单一真实或观测回归元 x^*，该 x^* 具有方差 $\sigma_{x^*}^2$，观测值 x 具有零均值的测量误差 v 以及有关的 σ_v^2。也就是说，回归为 $y=\beta x^*+u$，其中，$\text{E}[u|x^*]=0$，$\text{V}[u|x^*]=\sigma_u^2$，并且 $\text{Cov}[v,u]=0$，只是 x^* 在进行回归估计时，要用 x 观测变量代替。

在此情况下，式(26.7)被特别简化成：

$$\begin{aligned}\text{plim}\, \hat{\beta} &= \frac{\sigma_{x^*}^2}{\sigma_{x^*}^2+\sigma_v^2}\beta \\ &= \frac{1}{1+\sigma_v^2/\sigma_{x^*}^2}\beta \\ &= \beta[1-s/(1+s)]\end{aligned} \tag{26.8}$$

其中，$s=\sigma_v^2/\sigma_{x^*}^2$ 经常被称为信噪比[1](**noise-to-signal ratio**)，而将整个 $(1+s)^{-1}$ 称为可信率(**reliability ratio**)。从渐近形式上看，会向下偏倚起于 0，其偏倚程度直接依赖于信噪比。这种偏倚，还被称为衰减偏倚(**attenuation bias**)。该术语的含义非常直观，因为它表明研究者估计 x^* 变化对 y 的边际影响因测量误差而衰减。

---

〔1〕 又称为信号噪声比。——译者注

同样地，注意到：

$$V[y|x]=\sigma_u^2+\frac{\beta^2\sigma_v^2\sigma_{x^*}^2}{\sigma_{x^*}^2+\sigma_v^2}\geqslant\sigma_u^2$$

这蕴含着，测量误差不仅引起衰减偏倚，也会使方程误差方差变大。明确地讲，误差方差的减少将使方程的残差方差变小。

上面阐述了二变量回归包含截距项，会使截距的最小二乘估计量 $\bar{y}-\hat{\beta}\bar{x}$ 产生向上偏倚，其中，$(\bar{y},\bar{x})$ 都表示样本均值，$\bar{y}$ 与 $\bar{x}$ 是各自总体均值的一致估计值。克拉格(Cragg，1994)提出，运用"污染偏倚"(**contamination bias**)术语来表述测量误差对方程中另一个回归参数的这种效应。

举一个例子，考察小时工资对数对受教育年数的回归。假定受教育年数 x^* 测量时带有误差，并假定真实受教育年数的标准差为 2，而测量误差的标准差为 1，从而 $\sigma_{x^*}^2=4$，$\sigma_v^2=1$ 以及 $\sigma_x^2=5$。于是，$\text{plim}\,\hat{\beta}=0.8\times\beta$。比如，OLS 估计斜率系数为 0.04，意指实际上多受教育一年，则会使工资有 5%的而不是 4%的提高。

26.2.4 *推广*

把这种简单而优美的结果进行发展及推广，研究者经常会问，衰减偏倚是测量误差模型的一般特性吗？并且其衰减了多少？尽管该结果对更一般模型不一定成立，却提供了一种基准。豪斯曼(Hausman，2001)将由测量误差引起的衰减偏倚称为"经济计量学的铁律"。

若假定测量误差与真实不可观测值无关，则称此测量误差是"经典的"。虽然这样做方便，但该假设可能并不成立。实际上，在某些情况下，它不能成立。例如，当 x 是二值变量 0/1 时，测量误差将是一种分类误差。如果因错误分类而将 0 测量为 1，反之亦然，那么该种测量误差一定与真实值相关。

当存在一个以上回归元时，设 $\mathbf{X}^*=[\mathbf{x}^*\ \mathbf{Z}]$，并且如同上述情况一样，我们假定仅有一个回归元被观测到，且带有测量误差，即 $x=x^*+v$。于是，x 系数的最小二乘估计量表达式变成：

$$\text{plim}\,\hat{\beta}_{x|\mathbf{Z}}=\beta\left[1-\frac{\sigma_v^2}{\sigma_{x^*}^2(1-R_{x^*,\mathbf{Z}}^2)+\sigma_v^2}\right]\tag{26.9}$$

其中，$R_{x^*,\mathbf{Z}}^2$ 表示 $\mathbf{x}^*$ 对 $\mathbf{Z}$ 辅助回归的 R^2。倘若我们把 x^* 的方差重新解释成控制或去掉 $\mathbf{Z}$ 对 $\mathbf{x}^*$ 的线性影响之后方差，公式(26.9)本质上与式(26.8)[1]是一样的。最小二乘估计量的非一致性再次趋于 0，不过 β 的倍数小于单个回归元情况。不带测量误差的回归元系数也是非一致的，其方向依赖于 $\boldsymbol{\Sigma}_{\mathbf{x}^*\mathbf{x}^*}$[利瓦伊(Levi，1973)]。人们能将这种效应再次看成是污染偏倚。在这些特殊情况下，所阐述的衰减偏倚严重依赖于可加测量误差的假设。

当存在一个以上回归元时，进行测量并带有误差，就不可能利用非一致性方向的一般性结果。不过，在任何给定问题中，已知 $\boldsymbol{\Sigma}_{\mathbf{x}^*\mathbf{x}^*}$ 与 $\boldsymbol{\Sigma}_{\mathbf{vv}}$ 的知识，就能决定非一致

---

[1] 原著中这里为式(26.9)，是一个印刷错误，应为式(26.8)。——译者注

性方向。大部分研究都考察仅有一个回归元有测量误差,在此情况下,非一致性趋于 0。来自前面例子的直觉是,若不同回归元的测量误差是独立的,则每一种来源都将贡献给“自己”系数的偏倚,同时全部来源使得条件方差偏倚变大。克拉格(Cragg, 1994)分析了带有测量误差的多元回归模型,并证明偏倚之间的交互作用有不同来源。

26.2.5 线性面板模型测量误差

当运用面板数据时,回归元测量误差效应体现出一种混合形式。

假定混合面板模型 $y_{it}=\beta x_{it}^{*}+u_{it}$,其中可以观测到 $x_{it}=x_{it}^{*}+v_{it}$,而且为了简单起见,假定纯量回归元。如果我们估计单一横截面,那么上述结果仍成立。不过,当我们利用一年以上的个体数据进行估计时,就需要改动前面结果,因为回归元 x_{it}^{*} 更可能是正相关的,而不是对给定 i 时不同 t 而言是独立的。例如,若进行一阶差分,则得到回归:

$$\begin{aligned}\Delta y_{it}&=\beta\Delta x_{it}^{*}+\Delta u_{it}\\&=\beta\Delta x_{it}+\Delta u_{it}-\beta\Delta v_{it}\end{aligned}$$

(参见 21.6 节。)并定义 $\rho=\mathrm{Cor}[x_{it}^{*},x_{i,t-1}^{*}]$,那么:

$$\begin{aligned}\operatorname{plim}\hat{\beta}&=\beta+\left(\operatorname{plim}\frac{1}{N}\sum_{i=1}^{N}(\Delta x_{it})^{2}\right)^{-1}\left(\operatorname{plim}\frac{1}{N}\sum_{i=1}^{N}(\Delta x_{it}\Delta u_{it}-\beta\Delta x_{it}\Delta v_{it})\right)\\&=\beta-\frac{2\beta\sigma_{v}^{2}}{2(1-\rho)\sigma_{x^{*}}^{2}+2\sigma_{v}^{2}}\\&=\beta-\frac{\beta\sigma_{v}^{2}}{(1-\rho)\sigma_{x^{*}}^{2}+\sigma_{v}^{2}}\end{aligned}$$

这里用到了 $\mathrm{V}[\Delta v_{it}]=2\mathrm{V}[v_{it}]$与 $\mathrm{V}[\Delta x_{it}^{*}]=2(1-\rho)\mathrm{V}[x_{it}^{*}]$。

当 $\rho>0$ 时,此非一致性比横截面情况的要大一些。另外,当 $\rho\to 1$ 时,正如面板数据情形一样,非一致性变得相当大。通过运用差分法,这种非一致性得以减少,这里,$m>1$ 滞后除外,原因在于 $\mathrm{Cor}[x_{it}^{*},x_{i,t-m}^{*}]$ 关于 m 将是递减的。

26.3 识别策略

一般地讲,若没有额外假设,变量误差模型是不可识别的。可将这种陈述在二变量模型特殊情况下作如下解释。$\hat{\beta}$ 估计值,或更准确地讲,为 $\hat{\beta}$ 的概率极限,关于 β 与 s 信噪比的众多不同组合都是一致的。不过,如果能提供针对此问题的额外假设或信息,那么可能剔除基本参数的某些组合,基本参数与观测数据分布就是一致的。假如额外限制刚好足够获得唯一解,则称该模型是恰好识别的。假如额外限制足够多以致模型参数唯一识别,则称该模型是过度识别的。

测量误差模型的一般识别策略是,倘若没有进一步先验信息或数据,就要获得关注参数的界(**bounds**)不是点估计。如果有关于测量误差额外数据与/或信息,那么可使用额外的识别策略,诸如用工具变量估计,或通过矩约束识别。测量误差的

额外信息是一个宽泛的概念，它包括最古老的识别策略，以及将真实不可观测变量与其可观测部分联系起来的工具变量。比如，额外信息可产生衰减因子 $\sigma_{x^*}^2/(\sigma_{x^*}^2+\sigma_v^2)$ 的一致估计量，这使得调整偏倚的非一致性估计成为可能。最后，当有重复数据或核实数据[1]（**validation data**）可利用时，这些就能产生测量误差矩的有用信息。上述多种可能性，下面逐一加以分析。

26.3.1 *设置回归参数界*

重新考察 26.2 节的多元回归问题。已知该模型服从下述要求：方差 $\Sigma_{x^*x^*}$、Σ_{ww} 与 σ^2 必是半正定的。这与估计的正交性一起用于对系数必须位于的某个区域施加某种界。克莱珀和利默（Klepper and Leamer，1984）、万斯比克和梅杰（Wansbeek and Meijer，2000）都曾经研究某种一般性问题。一种更易于掌握的界方法特殊情况是下面将要阐述的逆向回归法。

逆向回归

在具有变量(y, x)的简单二变量回归模型中，正向回归（**direct regression**）意指 y 对 x 的回归，而逆向回归（**reverse regression**）意指 x 对 y 的回归。在具有 K 个协变量的一般多元回归情况下，仅仅存在一个正向回归，但有 K 个逆向回归。每一个逆向回归都有左边的错误测量内生变量，而其余内生变量与 y 位于右边。在带有测量误差的二变量回归情况下，容易证明，来自正向回归与逆向回归的估计斜率系数对真实斜率系数施加了其下界与上界。在分析测量误差效应时，这是一个潜在有用的结果。利默（Leamer，1978）对逆向回归给出了一个优秀的讨论。

首先，我们通过参照含有测量误差的简单二变量回归模型

$$\begin{aligned} y&=\beta x^*+u \\ x&=x^*+v \end{aligned} \tag{26.10}$$

考察逆向回归的逻辑性，其中，u 表示回归误差，v 表示解释观测变量 x 与无误差测量 x^* 之差，这里 x^* 可进入回归之中。我们假定 $u\sim\mathcal{N}[0,\sigma_u^2]$，并且 $v\sim\mathcal{N}[0,\sigma_v^2]$。

下述索拉里（Solari，1969）[以及利默（Leamer，1978）]的结构方法是，将 x^* 处理成似然函数中的未知参数。给定$(\mathbf{y},\mathbf{x})$数据，其联合似然是：

$$\begin{aligned} \mathrm{L}(\mathbf{x}^*,\beta,\sigma_u^2,\sigma_v^2) \propto (\sigma_u^2)^{-N/2}\exp\left[-\frac{1}{2\sigma_u^2}(\mathbf{y}-\beta\mathbf{x})'(\mathbf{y}-\beta\mathbf{x})\right] \\ \times(\sigma_v^2)^{-N/2}\exp\left[-\frac{1}{2\sigma_v^2}(\mathbf{x}^*-\mathbf{x})'(\mathbf{x}^*-\mathbf{x})\right] \end{aligned} \tag{26.11}$$

该条件在满足条件 $\sigma_u^2=0$ 与 $\mathbf{x}^*=\mathbf{x}$，或条件 $\sigma_v^2=0$ 与 $\mathbf{y}=\beta\mathbf{x}^*$ 的点上没有定义。当我们对这个服从约束的似然函数良好定义部分直接求极小值时，就得到二个纯量回归系数，一个是正向回归的 $\hat{\beta}_{\mathrm{D}}=\mathbf{y}'\mathbf{x}/\mathbf{x}'\mathbf{x}$，而另一个是逆向回归的 $\hat{\beta}_{\mathrm{R}}=\mathbf{y}'\mathbf{x}/\mathbf{y}'\mathbf{y}$。为了便于直观想象，注意到，如果 $\mathbf{x}$ 测量没有误差，那么 $\mathbf{y}$ 就是随机的，而 $\mathbf{x}$ 则不是，

[1] 又称为有效数据。——译者注

因此正向回归具有有意义的条件期望解释，从而只要 **x** 是随机的（测量时带有误差），条件期望 E[**x**|**y**]就有意义，因为这两个方程组被简化成 $x=(1/\beta)y-u/\beta+v$。也就是说，逆向回归会得出最小二乘估计值$(\widehat{1/\beta})$。可直接验证：

$$r_{xy}^2\hat{\beta}_{\mathrm{R}}=\hat{\beta}_{\mathrm{D}} \qquad (26.12)$$

$$\hat{\beta}_{\mathrm{D}}<\beta<\hat{\beta}_{\mathrm{R}}$$

其中，r_{xy}^2表示 x 与 y 之间样本相关性平方；其界表明 $\hat{\beta}_{\mathrm{D}}$ 是 β 的向下偏倚估计值，$\hat{\beta}_{\mathrm{R}}$ 是 β 的向上偏倚估计值。注意到，在运用微观经济数据时，这两个界可能非常广泛，其中，几乎总是有 $r_{xy}^2<0.5$ 的情况，甚至 $r_{xy}^2<0.1$ 更为普遍。

利默（Leamer，1978）曾经考察下述(y,x^*)模型，其中，(y,x^*)服从二变量正态分布，均值为$(\beta\bar{x}^*,\bar{x}^*)$，而协方差矩阵是：

$$\boldsymbol{\Sigma}=\begin{bmatrix}\sigma_u^2+\beta\sigma_{x^*}^2 & \beta\sigma_{x^*}^2\\ \beta\sigma_{x^*}^2 & \sigma_{x^*}^2+\sigma_v^2\end{bmatrix} \qquad (26.13)$$

利默（Leamer，1978，第 239～240 页）证明了，该模型的似然函数在正向回归估计量 $\hat{\beta}_{\mathrm{D}}$ 与逆向回归估计量 $\hat{\beta}_{\mathrm{R}}$ 之间的 β 任何值处都达到极大值。

前面分析表明，即使 β 是不可识别的，但对 β 值仍能施加一致界。这是界识别（**bounds identification**）的一种潜在有用的应用。该结果能以简单方式被推广到仅有一个回归测量时带有误差的多元回归情况。克莱珀和利默（Klepper and Leamer，1984）考察了对有 K 个回归元的多元回归的一种推广，那 K 个回归元测量时都带有误差。存在一个正向回归与 K 个逆向回归。在对每一个逆向回归估计之后，拟合回归得以重新正常化，对左边 y 而言具有单位系数。从而，$\hat{\boldsymbol{\beta}}_{\mathrm{D}}$ 是源自正向回归的估计向量。而$\hat{\boldsymbol{\beta}}_{\mathrm{R},j}(j=1,\cdots,K)$是来自第 j 个逆向回归的向量。借助于克莱珀和利默（Klepper and Leamer，1984）的结果，如果正向回归与逆向回归系数向量均位于同一个相限，那么 $\boldsymbol{\beta}$ 的可行值集合是正向回归与逆向回归的凸包，即 $\hat{\boldsymbol{\beta}}\in\{\hat{\boldsymbol{\beta}}|\hat{\boldsymbol{\beta}}=\lambda_{\mathrm{D}}\hat{\boldsymbol{\beta}}_{\mathrm{D}}+\lambda_1\hat{\boldsymbol{\beta}}_{\mathrm{R},1}+\cdots+\lambda_k\hat{\boldsymbol{\beta}}_{\mathrm{R},K}\}$，其中，$\lambda$ 权重是非负的且其和为 1。正向回归与逆向回归向量中最小系数是下界，而最大系数则是上界。倘若系数改变符号，则这些界将不存在。

除克莱珀和利默（Klepper and Leamer，1984）研究之外，在应用背景下，还有几种运用这些思想的研究。格林（Greene，1983）与戈德伯格（Goldberger，1984）将逆向回归用于测量纯量判别上。博林杰（Bollinger，2003）在工资与人力资本模型里分析了对黑人和白人之间的差距。博林杰（Bollinger，1996）曾经将界方法用于观测值类别被错误分类时类别虚拟变量回归的情况。

26.3.2 *利用工具变量进行识别*

解决识别问题的一种方法是，引入一个或多个矩约束，以此构成进一步识别信息。矩约束典型地表明，存在着一种工具变量，它与作为测量时带有误差的那个变量是相关的，或者从因果关系看是有关的。另外，这个变量与被建模结果的变量是无关的，或从因果关系上看无联系的。一旦把该种约束添加到最初模型上，从原则上看，这有助于解决识别问题。

从历史上看，工具变量估计量被用于线性模型的测量误差问题的潜在解决方法[雷厄瑟尔(Reiersøl, 1941)，德宾(Durbin, 1954)]。当右边一个或多个变量是内生的时候，从而与回归误差相关，工具变量方法具有类似动机。线性联立方程模型与线性测量误差模型均是同构的，因此，在测量误差背景下，运用工具变量类型估计量是自然而然的。

重新考察 4.8 节与 6.4 节的线性工具变量模型，其中，$\mathbf{y}=\mathbf{X}\beta+\mathbf{u}$ 且 $\mathrm{E}[\mathbf{u}|\mathbf{X}]\neq\mathbf{0}$，若可以利用 $\mathbf{Z}$ 的有效集合，当 $\dim[\mathbf{Z}]\geqslant\dim[\mathbf{X}]$时，就能使用 2SLS 估计量。

人们可使用回归元内生性的豪斯曼检验，对测量误差存在进行检验，参见 8.3 节。该检验的几种变形均可运用，并且 8.4 节已经给出一种变形检验。

实施工具变量估计量的主要问题在于寻找有效工具时出现的实际困难。一个好的工具拥有两个性质：其一，与方程误差零相关(一致性)，其二，与被替代成工具的那个变量高度相关(有效性)。典型地讲，这样的工具并不容易找到。尽管从理想形式上看，人们应从回归元与协变量之间的详细设定中明显地推导有效工具，但在实际应用时，利用特设方法是人们的共识。和完全系统设定方法不同，特设方法较为简单且不怎么使用。注意到，使工具拥有有效性的条件并不会产生自动挑选它们的方法。这些技术性条件能由某一个变量得以满足，该变量从因果关系上看与所研究的现象无关。人们考察与回归元强烈相关却与方程误差不相关的那种变量。文献中存在许多应用该种思想的有趣论文，例如，参见安格里斯特(Angrist, 1990)。假如出现选择情形，使用此类工具变量就可能引起争议及令人困惑。

我们考察收入对受教育的横截面回归例子中出现的几种可能工具。第一，若有双胞胎数据可以运用，则双胞胎受教育水平可用作工具变量，因为双胞胎受教育水平可能是相关的。于是，工具变量估计的一致性要求测量误差 v 与双胞胎受教育的任何测量误差之间没有相关性。第二，更一般地讲，与受教育相关的其他变量比如父母亲的教育水平或收入可用于工具变量。第三，进行调查时，可能要求受教育水平不止一个方程，或者如果数据来自面板研究时，受教育水平可利用其他年份的调查。这类工具变量可能与 x 高度相关，但在本例中，关于测量误差 x 与 z 之间不相关的假设就更难令人信服。

滞后变量经常用作工具变量，可是这些滞后变量同样拥有测量误差，因而只要测量误差的序列相关不是一个问题，该方法就会最小限度地得到满足。

在面板背景下，测量误差的效应可以很大。由于面板数据提供了多时期测量，所以一旦假定各个不同时期 x_{it}^* 有不相关的测量误差，工具变量估计就能用于给出一致参数估计值，参见萧政(Hsiao, 1986，第 63～65 页)。

26.3.3 *经由额外矩约束的识别*

有关方程与测量误差(u, v)的分布假设，能够确保识别。存在一种重要情况，即识别可借助错误测量变量的不可观测真实值的分布信息或假设。(y,x,x^*)的联合多元正态性假设，以及测量误差 v 与方程误差 u 分别服从 iid $\mathcal{N}[0,\sigma_v^2]$与 iid $\mathcal{N}[0,\sigma_u^2]$的假设，这些假设一起并不足以识别测量误差模型。不过，(x^*,u,v)的前四阶矩存在假设，同时每一个变量的三阶矩以及三阶交叉矩都是非零的假设一起

才能保证识别，而后一个假设表明违背了正态性，正如现在我们所要阐述的。

重新考察模型(26.10)：

$$y=\beta x^*+u$$
$$x=x^*+v$$

其简化式为 $y=\beta x+\varepsilon$，可通过工具变量方法加以估计，这里，$\varepsilon=u-\beta v$。不过，现在增加一个新信息：x^* 的分布在下面意义中不是正态的，即不论其偏度还是峰度均表现出非正态性。参见克拉格(Gragg, 1997)、达格奈斯和达格奈斯(Dagenais and Dagenais, 1997)、万斯比克和梅杰(Wansbeek and Meijer, 2000)。这些假设蕴含着下述 6 个条件：

$$\begin{aligned} &\mathrm{E}[(xy)x]=\beta\mathrm{E}[x^{*3}], && \mathrm{E}[(xy)u]=0 \\ &\mathrm{E}[(x^2)x]=\mathrm{E}[x^{*3}]+\mathrm{E}[(v^3)], && \mathrm{E}[(x^2)u]=-\beta\mathrm{E}[v^3] \\ &\mathrm{E}[(y^2)x]=\beta^2\mathrm{E}[x^{*3}], && \mathrm{E}[(y^2)u]=-\beta\mathrm{E}[(\varepsilon^3)] \end{aligned}$$

第一行蕴含，当 $\mathrm{E}[x_i^{*3}]\neq 0$ 时，积变量 x_iy_i 是一个有效工具。第二行蕴含，当 $\mathrm{E}[x_i^{*3}]\neq 0$ 但 $\mathrm{E}[v_i^3]=0$ 时，x_i^2 是一个有效工具；也就是说，x^* 是非正态的但 v 服从正态分布。实际上，出现偏度越大，则工具变量就越好。不过，由于 x^* 是不可观测的，所以关于 x^* 的任何推导都将需要建立在 x 基础上。最后一行蕴含，当 x^* 的三阶矩是非零的但 ε 的三阶矩却为 0 时，y_i^2 是一个有效工具。

已知这些矩条件，利用工具变量方法能一致估计出该模型参数。此例子阐明，在除 (y_i, x_i) 以外没有其他数据可利用时，额外矩假设如何帮助生成有用工具。

26.3.4 *重复数据*

如果要估计测量误差方差，那么可能有一种可供选择的求解方法。其基本思想如下，我们借助于某个依赖于测量误差的方差与协方差，能够调整回归元的样本二阶矩矩阵 $\mathbf{X}'\mathbf{X}$。注意，我们确实没有试图调整观测值本身。不过，可对样本矩加以调整，因为估计量是那些样本矩的函数。这种重要思想，也可被推广到更复杂模型上。

当已知测量误差方差 $\mathbf{\Sigma}_{\mathrm{vv}}$ 时，$\boldsymbol{\beta}$ 的一致估计值能利用

$$\tilde{\boldsymbol{\beta}}=(\mathbf{X}'\mathbf{X}-N\mathbf{\Sigma}_{\mathrm{vv}})^{-1}\mathbf{X}'\mathbf{y} \tag{26.14}$$

来获得，其中，N 表示样本量。该估计量是一致的，因为：

$$\begin{aligned} \tilde{\boldsymbol{\beta}} &= \mathrm{plim}(N^{-1}\mathbf{X}'\mathbf{X}-\mathbf{\Sigma}_{\mathrm{vv}})^{-1}\,\mathrm{plim}\,N^{-1}\mathbf{X}'\mathbf{y} \\ &= (\mathbf{\Sigma}_{\mathrm{x^*x^*}}+\mathbf{\Sigma}_{\mathrm{vv}}-\mathbf{\Sigma}_{\mathrm{vv}})^{-1}\mathbf{\Sigma}_{\mathrm{x^*x^*}}\boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned}$$

其中，$\mathrm{plim}\,N^{-1}\mathbf{X}'\mathbf{y}=\mathbf{\Sigma}_{\mathrm{x^*x^*}}\boldsymbol{\beta}$ 是利用 $\mathbf{X}=\mathbf{X}^*+\mathbf{V}$ 与 $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+(\mathbf{u}-\mathbf{V}\boldsymbol{\beta})$ 得到的。关于实际应用中估计 $\mathbf{\Sigma}_{\mathrm{vv}}$ 方法的详细考虑，参见克拉什斯基(Krashinsky, 2004)。

重复数据 (**data replication**)是下面一种情况：有不可观测 $\mathbf{X}^*$ 的无偏估计值可以利用。假定测量误差是可加的，并且有一个可观测的 $\mathbf{X}$：

$$\mathbf{X} = X^* + \mathbf{V}$$

若 $\mathbf{X}$ 是 $\mathbf{X}^*$ 的无偏估计值，则 $E[\mathbf{V}|\mathbf{X}^*]=\mathbf{0}$。当数据得以复制时，这直接意味着我们至少有 $\mathbf{X}$ 的两个测量值可以利用。它也意味着，运用多重测量值能获得 $\mathbf{V}$ 的矩估计，一旦假定多重测量的三则量误差都是不相关的。

假定有两个纯量(重复值)$X_{(1)}$ 与 $X_{(2)}$，使得 $X_{(j)}=X^*+V_{(j)}$，$i=1,2$。那么，$V[V_{(j)}]=E[X_{(j)}^2]-E[X_{(1)}X_{(2)}]$，该值可通过样本平均值 $N^{-1}\sum_i[X_{(j),i}^2-X_{(1),i}X_{(2),i}]$ 来估计。于是，运用式(26.14)估计回归参数。

例如，假定我们想要用高中 SAT 测试所取得的成绩预测大学一年级的年级平均分数(GPA)。众所周知，对 SAT 而言，观测分数会随着不同测试而变化。设 x^* 表示真实 SAT 分数，并设 x_1 与 x_2 表示两次不同 SAT 测试的观测 SAT 分数。于是，$x_1=x^*+v_1$，$x_1=x^*+v_2$，还假定 v_1 与 v_2 是独立的且有相等的方差。由此可得，$\mathrm{Cov}[x_1,x_2]=\sigma_{x^*}^2$，$V[x_1]=V[x_2]=\sigma_{x^*}^2+\sigma_v^2$，而且 $\mathrm{Cov}^2[x_1,x_2]=\sigma_{x^*}^2/(\sigma_{x^*}^2+\sigma_v^2)$。研究发现，该测试拥有可靠性 0.9，这意味着，从一次测验到另一次测验的相关性为 0.9，从而相关性平方为 0.81。因而，$\sigma_{x^*}^2/(\sigma_{x^*}^2+\sigma_v^2)=0.81$。由式(26.8)可得，$\mathrm{plim}\,\hat{\beta}=0.81\times\beta$，正因为测量误差的缘故，与普通最小二乘回归得出的结论相比，SAT 分数作为一年级大学 GPA 的更好预测公式。

26.3.5 *核实数据*

有时，核实样本也可能当作对最初响应的另一种检查。尽管核实样本(**validation sample**)属于关注总体，但它可能来自不同的独立来源。例如，病人对所受医疗服务问卷的回答，而服务提供者对核实调查做出反应。另一个例子是，雇员提供了一个事件的某种信息，该信息可由老板那里得到的同样信息来核实。经济学中的一个重要例子是，由邦德等人(Bound et al.，1994)进行的 PSID 核实研究。

设 $\mathbf{X}$ 表示回归元的观测值 $N\times K$ 阶矩阵，具有测量误差，并设 $\mathbf{X}_v$ 表示核实数据的 $M\times K$ 阶矩阵。我们借助 $\mathbf{X}_v$ 的列对 $\mathbf{X}$ 回归来运用核实数据，并生成预测值 $\mathbf{X}[\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{X}_v$，以此代替误差污染矩阵 $\mathbf{X}$。对于非线性模型，就要用更复杂的方法，参见李和塞帕斯基(Lee and Sepanski，1995)。

假如需使预测来自拟合良好的回归，那么将生成回归元代入关注回归之中的做法是一种实用的实际策略。生成回归元是真实值的估计值，从而受到估计不确定性的限制。就这点而论，在对回归系数样本方差进行估计时，应该将这种不确定性考虑进去。有关理论已在 6.8 节阐述。

26.4 非线性模型测量误差

正如上面清晰阐述的，非线性模型包含了大量的令人困惑的模型内容。要得到可应用于广泛模型的一般性结果，比如衰减偏倚，这是一项重要的挑战。并不令人感到惊讶的是，一般性结果都在简化假设下获得，而特定的结果则更多注重于特殊数据的复杂性及设定性。因此，文献中该专题的研究已经产生许多程序及方法，它们都是针对特殊模型加以考虑的，出现这种情形就不足为奇了。例如，在探索左

边有测量误差的二值结果模型时，自然关注错误分类问题，而在研究左边同样有测量误差的计数模型时，等价做法会关注报告不足或报告过度的问题。正是由于这类困难的推动，萧政（Hsiao，1992）将对一般模型求解的关注重点转向问题特定形式上。当涵盖特定模型结果时，会引发简单化与一般性结果直觉观点的丧失。因此，我们以某些选择的一般性结果开始讨论。

26.4.1 *通过工具变量识别*

线性变量误差模型的一般方法是工具变量法。对于（关于回归元）非线性回归模型，雨宫（Amemiya，1985）已经证明，工具变量估计量通常是非一致的，只有在使误差协方差矩阵变小的假设下，才会得出一致估计。

对前面提及观点的一种简单解释是，建立在回归方程

$$y=\beta_0+\beta_1 f(x^*)+\varepsilon \tag{26.15}$$

基础上，其中，$f(x^*)$是无误差纯量回归元 x^* 的光滑、可微以及有界的函数。观测变量 $x=x^*+v$，其中，v 表示测量误差。将 x^*代入，并利用 $f(x-v)$在 x 附近的泰勒级数展开式，得到：

$$y=\beta_0+\beta_1 f(x)+\varepsilon-\beta_1 f^{(1)}(x)v+\beta_1\sum_{j=2}^{\infty}f^{(j)}(x)(-v)^j/j! \tag{26.16}$$

其中，$f^{(j)}(\cdot)$表示 $f(\cdot)$的第 j 阶导数。考虑二次形式 $f(x)=x^2+\gamma x$，从而 $f^{(1)}(x)=2x+\gamma$，$f^{(2)}(x)=2$，而 $f^{(j)}(x)=0$，$j>2$。因而：

$$\begin{aligned}y&=\beta_0+\beta_1(x^2+\gamma x)+\varepsilon-\beta_1(2x+\gamma)v+\beta_1 2v^2/2\\&=\beta_0+\beta_1 x^2+\beta_1\gamma x+(\varepsilon-\beta_1 xuv-\beta_1\gamma v+\beta_1 v^2)\end{aligned} \tag{26.17}$$

所以，有效工具变量应与 x^2 及 x 相关，但与 $u=(\varepsilon-\beta_1 xv+\beta_1\gamma v+\beta_1 v^2)$不相关。很明显，$v$ 与 ε 各自都与工具无关，这还不够。这意味着，$f(x)$的工具变量要满足的条件比线性情况更为严格。

更一般地讲，使用泰勒级数加以近似，雨宫已经证明，非线性变量误差模型的工具变量并不会产生一致估计值，因为残差项既包括测量误差又包括观测误差污染变量。因此，不可能找到那种与观测变量高度相关并且与残差项不相关的工具变量。此外，从应用观点看，并不容易验证用于估计的工具变量的有效性，原因在于潜变量（x^*）与测量误差的信息有限。

26.4.2 *用重复数据识别*

当人们面临实施工具变量形式估计方法遇到困难时，存在两种可供选择的其他方法。第一种方法是，对给定观测值 x 时不可观测 x^* 的条件分布做出非常强的分布假设。这类假设可由其他技术性条件得以扩大，使得对模型参数识别成为可能。该方法遵照雨宫（Amemiya，1985）、萧政（Hsiao，1989）以及其他研究者的探索路线。

第二种方法是，考虑拥有每一个不可观测 x^*（记为 $x^{(j)}$）的大量测量值的可能

性。那么,每个 x^* 的重复测量的平均值代替不可观测回归元。由于当重复次数不断增大时,测量误差的协方差矩阵收缩到 0,所以得到非线性回归的一致估计,参见雨宫(Amemiya, 1985)。不幸的是,此类情况在经济计量学中极少遇到。

由于非线性测量误差模型确实不存在能用于识别与估计回归模型的共同结构信息,所以我们考察某些特定的非线性回归模型。

豪斯曼、纽韦和鲍威尔(Hausman, Newey, and Powell, 1995)分析了利用消费支出调查数据的多项式恩格尔曲线。他们所用的多项式关于参数为线性的。他们已经证明,在正则条件下,不论是工具变量还是额外测量,都能用于获得一致且服从渐近正态分布的估计值。在这个应用中,将邻近季度处理成重复的,并看成一个工具变量。他们进一步提出,通过多项式函数逼近一般非线性函数。不过,他们认为,在此情况下不能实施工具变量方法,从而需要真实回归元的另外测量。

李(Li, 2002)提出非线性变量误差问题的一般两阶段方法,该方法依赖于重复测量。在第一阶段,依据经验特征函数与傅里叶逆变换,可获得潜变量条件密度的非参数估计量。若运用此估计量,半参数非线性最小二乘估计量可借助于最小距离准则建立起来。他证明了,该估计量具有一致性。而且,该估计量在如下意义下是稳健的,即它不需要潜变量函数形式的任何知识。李方法能应用到任意非线性变量误差的情况,假如有重复测量可以利用。可是,该估计量的渐近分布尚未建立起来。

26.4.3 *因变量测量误差*

在线性回归模型中,因变量测量误差会使回归系数的标准误差变大,但不会导致估计量具有非一致性。在非线性模型中,同样情况则存在另一些后果。

一类应用问题是,考虑定性选择模型对响应的错误分类。这就产生了报告误差方面的文献。

离散选择模型

波特巴和萨默斯(Potorba and Summers, 1995)利用 CPS 数据研究了失业保险对失业持续期限的影响,将概率模型推广到可考虑劳动力市场状态过渡的错误分类情况上。特别地,他们关注三种类型:就业者、失业者以及非劳动力的潜在分类误差。他们探索了数据集合具有特定性质时的多项式 logit 模型:即假定所有个体在第一次调查中均被正确报告为失业者。他们结果表明,失业保险会使失业时期增大,同时对劳动力市场状态错误分类进行校正,会增强失业保险对持续期限长度影响的明显作用。不过,他们的模型建立在如下假设基础上,即假定报告误差的概率是固定的且与个体特征不相关,这一点正如作者承认的,"在实际应用中可能成立"。尽管作者声称:参数是一致的,但豪斯曼、阿布拉瓦亚和斯科特·莫顿(Hausman, Abrevaya, and Scott-Morton, 1998)已经论证,标准误差是非一致估计的,原因在于忽略掉估计误差概率的抽样变异性以及信息矩阵的非分块对角形式。

豪斯曼等人(Hausman et al., 1998)提出,估计带有错误分类的二值选择模型的一种参数方法。可是,他们的参数模型需要误差分布的知识。他们强调,若分布

不服从假定参数分布，则其参数估计可能是非一致的。

此外，他们引入了两阶段半参数方法。为了识别，该模型的关键性条件是，观测因变量的期望值是基本指标的增函数，他们证明，此条件比参数模型识别的条件更弱一些。与波特巴和萨默斯方法相比，他们的估计在错误分类概率是个体特征函数的意义下是稳健的。他们利用 CPS 与 PSID 证明，工作变动变量存在严重错误分类。

克莱因和舍曼(Klein and Sherman, 1997)针对潜在新录像产品的预计需求估计，发展出一种“轨道模型”(Orbit model)(具有有序选择模型与模型的特性)。他们发现，潜在消费者夸大了需求。该轨道模型是一种两阶段方法，其第一阶段估计实际未来需求的标准 Tobit 模型参数，而第二阶段则估计当前预计需求与实际未来需求之间的映射函数。此外，他们建立了轨道估计量的一致性与渐近正态性。不过，识别该模型需要下述假设：未来预计零需求将正好是零需求。这可能是一个强假设。

萧政和孙(Hsiao and Sun, 1999)运用先进电子设备需求方面的市场调查数据。他们证明，调查对象可能报告出有偏差的需求。他们提出一种随机化报告模型以及过高报告的单边响应偏倚模型，其中，不同参数概率被指派为真实选项或可供选择项(包括真实的)，对于真实显示性偏好来说，有 logit 或 probit 密度函数。他们发现，“数据存在大量的响应偏倚，若与那些调查对象真实表现其偏好所得出的估计值相比，修正的市场率及价格弹性似乎显得更有道理”。

计数回归

在非线性计数回归背景下，卡梅伦和特里维迪(Cameron and Trivedi, 1998)提出了，在可能未充分记录条件下，对计数数据进行建模的方法。该方法通过考虑二值记录结果产生了一种复合泊松模型与负二项计数模型。具体地讲，就事件的每一种单个结果而言，贝努利试验用于确定事件是否被记录。已知事件可能未被记录的概率为正的，则记录事件分布的均值与方差均会小于实际事件的分布情况。他们进一步用似然估计、准广义伪极大似然法以及基于矩方法研究了模型估计。他们运用蒙特卡罗方法进行研究发现，信用极大似然估计量的效果在样本量为 50 或更多时表现良好。

乔丹等人(Jordan et al., 1997)曾经给出泊松回归模型中误差变量的一个应用。在对日本 5 个省份因胃癌死亡的研究中，他们注意到，协变量(比如血浆番茄红素水平)是未知的，并通过随机选择的全体人员加以估计，因而受限于抽样误差。运用测量误差服从正态分布的假设，他们借助于吉布斯抽样获得参数的后验分布，再据此实施贝叶斯方法。其结果显示，当最初样本很小时，修正模型会得出参数更准确的估计值。

26.4.4 *含有协变量测量误差的泊松回归*

现在，我们以更详细的方式考察非线性回归模型含有协变量可加测量误差的一个特定例子。用该例子既可阐明此类测量误差的后果，又可说明可行估计策略。

郭和李(Guo and Li, 2002)已经证明，协变量测量误差一般会导致观测数据出

现过度分散。他们还运用蒙特卡罗模型证明，倘若由测量误差引起的过度分散没有被正确建模成因不可观测异质性而导致的过度分散，则将出现偏倚。因此，人们不应因有过度分散就得出下述结论，即保证模型具有不可观测异质性。

斯特凡斯基(Stefanski, 1989)与中村(Nakamura, 1990)都曾经提出一种修正得分估计量(**corrected score estimator**)，当存在测量误差时，该估计量是一致的。尤其是，中村(Nakamura, 1990)给出了当测量误差服从正态分布，并且还可利用重复数据时，修正得分函数的一种闭形式。与之相比，郭和李(Guo and Li, 2002)则推广了中村方法。

测量误差与过度分散

本节考察泊松回归模型，其中，离散随机变量 y 服从泊松分布，该分布参数$\mu=\exp(\mathbf{x}^{*\prime}\boldsymbol{\beta})$，$\boldsymbol{\beta}$ 表示 $K\times1$ 维参数。众所周知，泊松回归模型具有等分散性质，即：

$$\mathrm{E}[y|\mathbf{x}^*]=\mathrm{V}[y|\mathbf{x}^*] \tag{26.18}$$

若测量误差为可加的，则：

$$\mathbf{x}=\mathbf{x}^*+\varepsilon$$

其中，ε 被假定为与不可观测潜变量 $\mathbf{x}^*$ 是独立的，ε 的均值为 0 且方差协方差矩阵是 Σ_ε。此符号涵盖了所有解释变量或部分解释变量测量时具有误差。

测量误差会增大分散性[切舍(Chesher, 1991)]。这适用于下述意义上的泊松回归，即虽然式(26.18)对于给定 $\mathbf{x}^*$ 时 y 的条件均值与方差来说成立，但以 $\mathbf{x}$ 为条件却改变了结果。相反，我们得到，$\mathrm{E}[y|\mathbf{x}]<\mathrm{V}[y|\mathbf{x}]$，部分原因在于 $\mathrm{E}[y|\mathbf{x}^*]\neq\mathrm{E}[y|\mathbf{x}]$，并且 $\mathrm{V}[y|\mathbf{x}^*]\neq\mathrm{V}[y|\mathbf{x}]$。

如果用 $g(\mathbf{x}^*|\mathbf{x})$表示给定 $\mathbf{x}$ 时 $\mathbf{x}^*$ 的条件密度，那么郭和李已经证明：

$$\begin{aligned}\mathrm{E}[y|\mathbf{x}] &= \int\mathrm{E}[y|\mathbf{x}^*]g(\mathbf{x}^*|\mathbf{x})d\mathbf{x}^* \\ &= \int\mathrm{E}[y^2|\mathbf{x}^*]g(\mathbf{x}^*|\mathbf{x})d\mathbf{x}^*-\int(\mathrm{E}[y|\mathbf{x}^*])^2g(\mathbf{x}^*|\mathbf{x})d\mathbf{x}^*\end{aligned} \tag{26.19}$$

并且使用式(26.18)，给定 $\mathbf{x}$ 时 y 的条件方差是：

$$\mathrm{V}[y|\mathbf{x}]=\int\mathrm{E}[y^2|\mathbf{x}^*]g(\mathbf{x}^*|\mathbf{x})d\mathbf{x}^*-\left[\int\mathrm{E}[y|\mathbf{x}^*]g(\mathbf{x}^*|\mathbf{x})d\mathbf{x}^*\right]^2 \tag{26.20}$$

通过比较式(26.19)与式(26.20)可以发现，式(26.19)括号内的第一项与式(26.20)第一项是一样的。利用这一点，郭和李曾经证明：

$$\left[\int\mathrm{E}[y|\mathbf{x}^*]g(\mathbf{x}^*|\mathbf{x})d\mathbf{x}^*\right]^2\leqslant\int(\mathrm{E}[y\mid\mathbf{x}^*])^2g(\mathbf{x}^*|\mathbf{x})d\mathbf{x}^* \tag{26.21}$$

该式被解释成测量误差导致了过度分散。

测量误差模型的估计

当 $\mathbf{x}$ 被测量误差所污染时，基于可观测值$(y,\mathbf{x})$的极大似然估计或非线性最小二乘估计并没有给出其一致估计。当协变量 $\mathbf{x}^*$ 被 $\mathbf{x}$ 代替时，则称为“朴素”模型。

这种考虑存在两个问题。第一，当存在测量误差时，为什么用极大似然法得出

非一致估计？第二，会有一致估计吗？假如我们采用遵照斯特凡斯基（Stefanski，1989）与中村（Nakamura，1990）的广义线性模型的修正得分估计方法，则对第二个问题的回答为“有一致估计”。

支撑修正得分估计量的思想是，给定真实自变量 $\mathbf{x}^*$ 与应变量 y，关于 $\mathbf{x}$ 修正估计的条件分布是以极大似然估计值为中心的，这提供了关注参数真实值的一致估计。

非一致估计量与一致估计量

假定 N 个观测值 $(y_i, \mathbf{x}_i^*)$，$i=1,\cdots,N$ 均来自泊松分布，其概率质量函数为：

$$\Pr[Y_i=y_i\,|\,\mathbf{x}_i^*]=\frac{e^{-\mu_i(\boldsymbol{\beta}_0)}\mu_i(\boldsymbol{\beta}_0)^{y_i}}{y_i!}$$

其中，$\mu_i(\boldsymbol{\beta}_0)=\exp(\mathbf{x}_i^{*\prime}\boldsymbol{\beta}_0)$。已知观测值 $(y_i,\mathbf{x}_i^*)$，$i=1,2,\cdots,N$，由于平均对数似然函数的概率极限：

$$\begin{aligned}\operatorname{plim} N^{-1}\ln \mathrm{L}(\boldsymbol{\beta})&=N^{-1}\sum_i\{-e^{\mathbf{x}_i^{*\prime}\boldsymbol{\beta}}+y_i\mathbf{x}_i^{*\prime}\boldsymbol{\beta}-\ln y_i!\}\\&=\mathrm{E}_{y,\mathbf{x}^*}[-e^{\mathbf{x}^{*\prime}\boldsymbol{\beta}}+y\mathbf{x}^{*\prime}\boldsymbol{\beta}-\ln y_i!]\end{aligned}\tag{26.22}$$

在 $\boldsymbol{\beta}=\boldsymbol{\beta}_0$ 处取极大值，故极大似然估计量 $\hat{\boldsymbol{\beta}}$ 是一致的。

假定我们观测到 $\mathbf{x}_i$ 而不是 $\mathbf{x}_i^*$，其中，$\mathbf{x}_i=\mathbf{x}_i^*+\varepsilon_i$，$\varepsilon_i\sim\mathcal{N}[\mathbf{0},\ \boldsymbol{\Sigma}_\varepsilon]$，$\varepsilon_i$ 与 $\mathbf{x}_i^*$ 是独立的。那么，$y_i\,|\,\mathbf{x}_i$ 并不服从泊松分布。尽管如此，若人们使用“朴素泊松模型”，则所得到的估计量 $\tilde{\boldsymbol{\beta}}$ 使：

$$Q(\boldsymbol{\beta})=N^{-1}\sum_i\{-e^{\mathbf{x}_i'\boldsymbol{\beta}}+y_i\mathbf{x}_i'\boldsymbol{\beta}-\ln y_i!\}\tag{26.23}$$

达到极大值。这种错误设定对数似然函数收敛到：

$$\operatorname{plim} Q(\boldsymbol{\beta})=\mathrm{E}_{y,\mathbf{x}^*}[-e^{\mathbf{x}^{*\prime}\boldsymbol{\beta}}+y\mathbf{x}^{*\prime}\boldsymbol{\beta}-\ln y_i!]+\mathrm{E}_{\mathbf{x}^*}[-e^{\mathbf{x}^{*\prime}\boldsymbol{\beta}}](\mathrm{E}_\varepsilon[e^{\varepsilon'\boldsymbol{\beta}}]-1)\tag{26.24}$$

一般地讲，它没有在 $\boldsymbol{\beta}=\boldsymbol{\beta}_0$ 处取极大值。因而，$\tilde{\boldsymbol{\beta}}$ 关于 $\boldsymbol{\beta}_0$ 是非一致的。

一旦对目标函数进行适当修正，则会得出一致估计量。式(26.22)与式(26.24)蕴含：

$$\{\operatorname{plim} Q(\boldsymbol{\beta})-\mathrm{E}_{\mathbf{x}^*}[-e^{\mathbf{x}^{*\prime}\boldsymbol{\beta}}](\mathrm{E}_\varepsilon[e^{\varepsilon'\boldsymbol{\beta}}]-1)=\operatorname{plim} N^{-1}\ln \mathrm{L}(\boldsymbol{\beta})$$

这建议，对目标函数

$$Q^+(\boldsymbol{\beta})=N^{-1}\sum_i\{-e^{\mathbf{x}_i'\boldsymbol{\beta}}+y_i\mathbf{x}_i'\boldsymbol{\beta}-\ln y_i!\}-\mathrm{E}_{\mathbf{x}^*}[-e^{\mathbf{x}^{*\prime}\boldsymbol{\beta}}](\mathrm{E}_\varepsilon[e^{\varepsilon'\boldsymbol{\beta}}]-1)$$

求极大值，因为 $Q^+(\boldsymbol{\beta})$ 收敛到 $\operatorname{plim} N^{-1}\ln \mathrm{L}(\boldsymbol{\beta})$。现在，已知 $\mathbf{x}^*$ 与 ε 是独立的，则：

$$\mathrm{E}_{\mathbf{x}^*}[-e^{\mathbf{x}^{*\prime}\boldsymbol{\beta}}]\mathrm{E}_\varepsilon[e^{\varepsilon'\boldsymbol{\beta}}]=\mathrm{E}_{\mathbf{x}^*,\varepsilon}[-e^{(\mathbf{x}^*+\varepsilon)'\boldsymbol{\beta}}]=-\mathrm{E}_{\mathbf{x}}[e^{\mathbf{x}'\boldsymbol{\beta}}]$$

它可通过 $-N^{-1}\sum_i e^{x_i'\beta}$ 得到一致估计。经过某些简化运算，可得到对 $Q^+(\boldsymbol{\beta})$ 求极大值，它等价于对

$$Q^{++}(\boldsymbol{\beta})=N^{-1}\sum_i\{y_i\mathbf{x}_i'\boldsymbol{\beta}-\ln y_i!\}-\mathrm{E}_{\mathbf{x}^*}[e^{\mathbf{x}^{*\prime}\boldsymbol{\beta}}]\tag{26.25}$$

求极大值。这就得出 $\boldsymbol{\beta}_0$ 的一致估计。当具体求解时,需要 $\mathrm{E}_{\mathbf{x}^*}[e^{\mathbf{x}^{*\prime}\boldsymbol{\beta}}]$的合适估计值,但若可利用重复数据,这样做就行得通。如果对解释变量分布加以设定,该设定至多相差一个未知参数,那么这些未知参数能用重复测量值加以估计。因此,就能得到 $\mathrm{E}_{\mathbf{x}^*}[e^{\mathbf{x}^{*\prime}\boldsymbol{\beta}}]$的估计值。

对式(26.25)求极大值的估计量 $\hat{\boldsymbol{\beta}}_C$,被郭和李(Guo and Li, 2002)称为修正得分估计量(**corrected score estimator**),因为它是修正得分函数 $\sum_i(y_i\mathbf{x}_i-\mathrm{E}_{\mathbf{x}^*}[\mathbf{x}^*e^{\mathbf{x}^{*\prime}\boldsymbol{\beta}}]=\mathbf{0}$ 的平方根。郭和李也建立了该估计量的渐近正态性。估计渐近协方差矩阵 $\hat{\mathrm{V}}[\hat{\boldsymbol{\beta}}_C]=N^{-1}\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}$,其中:

$$\hat{\mathbf{A}}=\mathrm{E}_{\mathbf{x}^*}[e^{\mathbf{x}^{*\prime}\hat{\boldsymbol{\beta}}_C}\mathbf{x}^*\mathbf{x}^{*\prime}]$$

$$\hat{\mathbf{B}}=N^{-1}\sum_i(y_i\mathbf{x}_i-\mathrm{E}_{\mathbf{x}^*}[e^{\mathbf{x}^{*\prime}\hat{\boldsymbol{\beta}}_C}\mathbf{x}^*])(y_i\mathbf{x}_i-\mathrm{E}_{\mathbf{x}^*}[e^{\mathbf{x}^{*\prime}\hat{\boldsymbol{\beta}}_C}\mathbf{x}^*])'$$

中村(Nakamura, 1990)做出过一个较强假设:测量误差 ε 服从正态分布 $\mathcal{N}[\mathbf{0},\boldsymbol{\Omega}]$。于是:

$$\exp(\mathbf{x}^{*\prime}\boldsymbol{\beta})=\mathrm{E}_{\mathbf{x}|\mathbf{x}^*}[\exp(\mathbf{x}'\boldsymbol{\beta}-(\boldsymbol{\beta}'\boldsymbol{\Omega}\boldsymbol{\beta}/2))]$$

应用期望迭代律:

$$\mathrm{E}_{\mathbf{x}^*}[\exp(\mathbf{x}^{*\prime}\boldsymbol{\beta})]=\mathrm{E}_{\mathbf{x}}[\exp(\mathbf{x}'\boldsymbol{\beta}-(\boldsymbol{\beta}'\boldsymbol{\Omega}\boldsymbol{\beta}/2))]$$

它能通过 $N^{-1}\sum_i[\exp(\mathbf{x}_i'\boldsymbol{\beta}-(\boldsymbol{\beta}'\boldsymbol{\Omega}\boldsymbol{\beta}/2))]$得到一致估计。因而,对于式(26.13)中的 $Q(\boldsymbol{\beta})$来说,式(26.14)给出的概率极限可简化成:

$$\mathrm{plim}\,Q(\boldsymbol{\beta})=N^{-1}\sum_i[y_i\mathbf{x}_i'\boldsymbol{\beta}-\ln y_i!-\exp(\mathbf{x}_i'\boldsymbol{\beta}-(\boldsymbol{\beta}'\boldsymbol{\Omega}\boldsymbol{\beta}/2))]$$

这是中村(Nakamura, 1990)曾经给出的修正对数似然函数。对 $\boldsymbol{\beta}$ 求极大值,会得到 $\boldsymbol{\beta}_0$ 的一致估计值。

中村方法提醒人们注意,当已知测量误差协方差矩阵的估计值时,对含有测量误差线性回归[参见式(26.14)]进行估计的一种方法。如同那种情况一样,为了获得中村修正对数似然函数的极大值,人们需要知道 $\boldsymbol{\Omega}$ 的知识,即测量误差的协方差矩阵。这可由重复数据得出。不过,如果协方差在多数情况下为离散的,那么测量误差正态性就是一个不切合实际的假设。在此情况下,郭和李的估计量更引人注目。

对于多变量 $\mathbf{x}^*$ 的情况,即使已知 $\mathbf{x}^*$ 的分布,也不能直接计算出 $\mathrm{E}[\exp(\mathbf{x}^{*\prime}\boldsymbol{\beta})]$,因为此时涉及多重积分。而基于模拟的方法[李(Li, 2002)]提供了该问题的一种可行方法。

其他几种非线性变量误差模型的具体求解也需要重复观测值,例如,参见萧政(Hsiao, 1992),豪斯曼、纽韦和鲍威尔(Hausman, Newey, and Powell, 1995)。面板数据在个体水平上提供了重复观测值。例如,考察下面纯量回归元 x^* 的情形:有两个 x 重复可利用,因为 $x_{ij}=x_i+\varepsilon_{ij}$,对于 $i=1,\cdots,N$, $j=1,2$。于是,σ_ε^2 的基于矩的一致估计量是 $\hat{\sigma}_\varepsilon^2=\sum_i(x_{i1}^2+x_{i2}^2-2x_{i1}x_{i2})/2N$。因而,不论是 $\mathbf{x}^*$ 的均值还是 $\mathbf{x}^*$ 的方差,均能被估计出来。

26.5 衰减偏倚模拟例子

线性模型的解析结果已由26.2节给出，但要获得非线性模型的结果就相当困难。这里，我们给出两个模拟例子，一个是logit模型，另一个是关于log为线性的模型，以阐明含有回归元测量误差的非线性回归的衰减偏倚。

在第一个例子中，数据生成过程是满足下述条件的logit模型：

$$y^* = \alpha^* + \beta^* x^* + \varepsilon$$
$$x^* \sim \mathcal{U}[1,2], \ \varepsilon \sim \text{逻辑斯蒂分布}$$
$$y = \begin{cases} 0, & \text{当 } y^* \leqslant 0 \\ 1, & \text{当 } y^* > 0 \end{cases}$$

其复杂情况是，x^*测量时有误差，从而：

$$x = x^* + v$$
$$v \sim \mathcal{N}[0, \sigma_v^2]$$

由于$x^* \sim \mathcal{U}[0,1]$，其方差为$\sigma_{x^*}^2 = 1/12$，信噪比是$s = 12\sigma_v^2$。人们可以估计$y$对$x$的logit回归；而不是$y$对$x^*$的logit回归。

为了实施模拟演算，我们完成y对x的logit回归，使用包括0的6个不同信噪比，以此作为标准衡量评估该模型。样本量被固定在1 000，并使用100个模拟重复值。

表26.1给出100个重复值的$(\hat{\alpha}, \hat{\beta})$平均值，其中，$\hat{\alpha}$与$\tilde{\beta}$是来自$y$对$x$的logit回归，而不是$y$对$x^*$的正确logit回归，估计截距与斜率，对于样本$N=1\ 000$是就$\sigma_v^2$的6个不同值而言得到6个不同信噪比$s$。用满足$s=0$的第一列衡量评估该模型。回顾同样背景下普通最小二乘线性回归斜率系数方面的乘法偏倚分别是$1/(1+s)$或0.96、0.8、0.5、0.2以及0.1。此处，偏倚有类似方向，只是对logit回归而言，这些偏倚显得较大。

表26.1 含有测量误差的logit回归衰减偏倚

噪声/信号	0	0.04	0.25	1	4	9
平均$\hat{\alpha}$	0.785	1.062	1.406	1.548	1.570	1.596
平均$\hat{\beta}$	1.799	1.224	0.446	0.125	0.037	0.012

第二个例子是一个二变量关于对数为线性的乘法模型，其中，$\alpha=2$，$\beta=0.4$，两个变量都含有可加测量误差。在这种情况下，其设置如下：

$$y^* = 4x^{*0.4}u, \ u \sim \mathcal{N}[10, 0.000\ 1]$$
$$x^* = 100 + \mathcal{U}[0,1]$$
$$y = y^* + \varepsilon_y, \ \varepsilon_y \sim \mathcal{N}[0, \sigma_y^2]$$
$$x = x^* + \varepsilon_x, \ \varepsilon_x \sim \mathcal{N}[0, \sigma_x^2]$$

就模拟而言，样本量为 1 000，而复制次数为 100 次。当实验不断进行时，我们改变 x^* 方差值，从而得到下面 $\sigma_x^2/\sigma_{x^*}^2$ 的一些值：0. 001，0. 01，0. 1，1，5，10，50，100，1 000 和 5 000。

表 26. 2 上面一行给出各种不同实验斜率系数的平均值，其信噪比也不断变化。这再次表明，衰减偏倚非常明显。

表 26. 2　含有可加测量误差非线性回归的衰减偏倚

$\sigma_x^2/\sigma_{x^*}^2$	0. 000 25	0. 002 5	0. 025	0. 25	2. 5	25
平均 $\hat{\beta}$	0. 393	0. 383	0. 341	0. 217	0. 063	0. 020

这两个例子所得到的结果与支撑"经济计量学铁律"的假设相一致。

26. 6　文献注释

迄今为止，万斯比克和梅杰（Wansbeek and Meijer, 2000）的书是从经济计量学的视角撰写的关于测量误差方面的最新的和最综合的著作。从深度上看，该书涵盖本章绝大多数专题，尤为强调线性模型。作者还在该书中提供几章将测量误差与因子模型、潜变量模型、结构方程模型联系的内容。在讨论结果时，作者避免"可以证明"之类的术语支持他们的详细推导。另外，豪斯曼（Hausman, 2000）也从经济计量学视角给出他与其同事研究获得最新结果的一个综述。邦德、布朗和马蒂威茨（Bound, Brown, and Mathiowetz, 2001）针对劳动力市场的测量误差问题做了一个综述。

从统计文献上看，已经很好地建立测量误差专题。富勒（Fuller, 1987）的书是一个极为有用的参考文献；尤其是，可以看到，当已知信噪比时，他对可用于该问题的正交回归方法进行的研究。尽管本章给出的线性模型是经济计量文献中非常标准的内容，但读者也应注意到另一种伯克森误差模型（**Berkson error model**），其中，不可观测真实变量被假定成常值，只是不完美测量变量受限于误差，而安格里斯特和克鲁格（Angrist and Krueger, 1999）对非经典测量误差（**nonclassical measurement error**）模型进行了讨论。马丹斯基（Madansky, 1959）给出了早期数值结果与方法。也可参见斯特凡斯基（Stefanski, 2000）。

26. 2　比约恩（Biorn, 1992）曾经分析了含有测量误差的面板数据模型。

26. 3　戈德伯格（Goldberger, 1984）与格林（Greene, 1983）在对康韦和罗伯茨（Conway and Roberts, 1983）的评注中，分析了有趣的逆向回归。利默（Leamer, 1978）已经从贝叶斯观点中提供逆向回归的富有深刻见解的讨论。哈恩和豪斯曼（Hahn and Hausman, 2002）运用逆向回归观点建立了关于测量误差问题工具变量方法有效性的设定检验。其关切内容是，可利用工具可能是弱的，得出不好的估计。哈恩和豪斯曼思想是完成正向回归的工具变量估计，这里的错误测量变量出现在方程右边。逆向回归在左边具有相同的错误测量变量。这种回归通过将相同工具变量作为正向回归，借助于工具变量也能得以估计。

26. 4　非线性模型的测量误差文献显得更为散乱。对经济计量学家来说，雨

宫(Amemiya, 1985)的书尤其有用。从统计观点看,卡罗尔等人(Carroll et al., 1995)考察了非线性模型,特别是广义线性模型,回归元含有可加的测量误差,所用的一系列方法包括有重复数据可利用时的那些方法。李、特里维迪和郭(Li, Trivedi, and Guo, 2003)发展并应用了一种测量误差变量模型,其中,计数响应变量具有测量误差。

习 题

26-1 考察二变量误差模型斜率参数的衰减偏倚结果[26.2.3 节的式(26.9)]。将该模型推广到含有截距项的模型上。

(a) 推导类似的截距项的测量误差偏倚结果。

(b) 推导类似的关于最小二乘截距估计的界识别,这类似于 26.3.1 节的式(26.12)。

26-2 [改编自博林杰(Bollinger, 2003)。]考察如下形式的一种多元回归模型,其中,纯量回归元 x 测量时有误差,而其他回归元 $\mathbf{z}$ 向量则没有测量误差。

(a) 维持二变量误差模型的测量误差假设,将衰减偏倚结果与界识别结果推广到本题情况。

(b) 检验对二变量情况进行专门化研究的那些新结果。

26-3 [改编自万斯比克和梅杰(Wansbeek and Meijer, 2000)。]考察二次型回归模型 $y=\alpha+\beta x^*+\gamma x^{*2}+\varepsilon$,其中,回归元 $x^*=x+v$ 为可观测的,而 v 为测量误差。假定(x^*,ε,v)是互不相关的且服从正态分布,同时所有变量均值为 0。

(a) 比较 β 与 γ 的最小二乘估计量偏倚。

(b) 该模型是可识别的吗?将最后结果与来自二变量线性变量误差模型的结果加以比较。

26-4 代际流动能力文献使用了下述模型[索伦(Solon, 1992);齐默尔曼(Zimmerman, 1992)]:

$$Y_i^{儿子}=\alpha+\beta Y_i^{父亲}+\varepsilon_i^{儿子} \tag{26.26}$$

其中,$\varepsilon_i \sim \text{iid } \mathcal{N}[0,\sigma^2]$。这里,$Y$ 表示持久地位(诸如持久收入)的测量,β 测量回归接近于经济地位平均水平的程度。假定不可以观测持久地位。可是,当前状况 Y_{it} 是可观测的,并满足 $Y_{it}=Y_i+\gamma X_{it}+w_{it}$,因此,$Y_{it}$ 是由称为持久地位的个体固定效应 Y_i 系统因素 X_{it} 与暂时误差成分 w_{it} 所构成的。设 $\hat{\gamma}$ 表示最小二乘系数拟合,并设:

$$Y_{it}-\hat{\gamma}X_{it}=Y_i+(\gamma-\hat{\gamma})X_{it}+w_{it}=Y_i+v_{it}$$

(a) 设 $\bar{Y}_i^{父亲}=T^{-1}\sum_{t=1}^{T}Y_{it}^{父亲}$ 表示父亲地位平均值,用它作为自变量,以此作为式(26.26)中的不可观测持久地位。设 $\hat{\beta}_{\text{avg}}$ 表示相应回归系数。证明:$\text{plim}\,\hat{\beta}_{\text{avg}}=\beta P_Y$,其中,$P_Y=\sigma_Y^2/(\sigma_Y^2+T^{-1}\sigma_\varepsilon^2)$

(b) 假定父亲收入的暂时成分遵从下面自回归模型,$v_{it}^{父亲}=\rho v_i^{父亲}+\xi_{it}$,其中,$\xi_i\sim\mathcal{N}[0,\sigma_\xi^2]$,$i=1,\cdots,T$。证明:现在 $\text{plim}\,\hat{\beta}_{\text{avg}}=\beta P_Y^*$,其中,$\beta P_Y^*=\sigma_Y^2/(\sigma_Y^2+T^{-1}V)$,而 $V=\sigma_\xi^2[T(1-\rho^2)]^{-1}[(1+2\rho)\{T-(1-\rho^T)/(1-\rho)\}/T(1-\rho)]$。

27

缺失数据与估算

27.1 引 论

调查数据出现缺失现象是因调查问题无回答或部分回答而引起的一个古老问题。无回答的理由包括:不愿意提供所问信息、很难回忆起过去发生的事件、不知道正确的回答。估算[1](**imputation**)是一种估计或预测缺失观测值的过程。

在本章,我们研究含有数据向量的回归背景,这里的数据向量为$(y_i, \mathbf{x}_i)$,$i=1,\cdots,N$。对于某些观测值来说,$\mathbf{x}_i$ 的某些元素或$(y_i, \mathbf{x}_i)$元素之中的某些元素出现缺失。因而,需要考虑一系列问题。什么时候我们应着手分析仅有完整观测值?什么时候应试图填上由缺失观测值而引起的缺口?什么样的估算方法可以利用?一旦获得缺失观测值的估算,又怎样进行估计与推断?

假如数据集出现缺失观测值,而且这些缺口能利用统计上合理的方法加以填补,则这样做的益处源于拥有更大的且可能更有代表性的样本,并在理想环境下可实施更准确的推断。估计缺失数据的成本来自做出支撑生成缺失观测值代表性方法的(可能错误)假设,并且来自任何这种方法固有的近似误差。另外,在用估算值代替缺失数据之后,由数据扩充而引发的统计推断会更加复杂,因为此类推断必须考虑到因估算而引入的近似误差。

作为调查无回答与因一组调查对象损耗而出现数据缺口的情况经常发生。对缺失值估算可能由官方机构来完成,以此生成与维护公用调查数据库,或者由那些使用数据建模者完成。在前者情况下,官方机构可以拥有更广泛的信息,包括机密信息/秘密信息,这些信息能在估算过程中得到利用。在后者情况下,建模者具有特定的建模框架,在估算过程时,则要利用这种建模框架。

一个有趣的缺失数据例子是,在消费者财政调查背景下(Survey of Consumer Finances)[肯尼克尔(Kennickell, 1998)]出现的问题。因为消费者财政问题极为敏感,所以调查表出现收入与财富信息的大量缺口,美国联邦储备的分析人员针对连续变量与离散变量,发展并实施一些复杂估算算法,既利用公开可用的收入与财富的调查信息,又有来自人口普查数据的保密信息。

---

〔1〕 又称为借补、设算。——译者注

图 27.1 给出回归元出现缺失数据的某些潜在模式。某一个数据集具有一个纯量因变量 y 以及三个回归元 x_1、x_2、x_3，它们中每一个都有观测值，那么将它们叠放成 $(\mathbf{y}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$。在 A 组调查对象中，都是完整数据，但观测值 $\mathbf{x}_1$ 有一些缺失，B 组调查 $(\mathbf{y}, \mathbf{x}_3)$ 是完整数据，而数据 $(\mathbf{x}_1, \mathbf{x}_2)$ 出现缺失值，使得 $\mathbf{x}_1$ 与 $\mathbf{x}_2$ 永远不会同时被观测到。C 组调查是全部三个回归元都出现缺失观测值时的一般缺失观测值模式，但不存在特定的缺失模式。

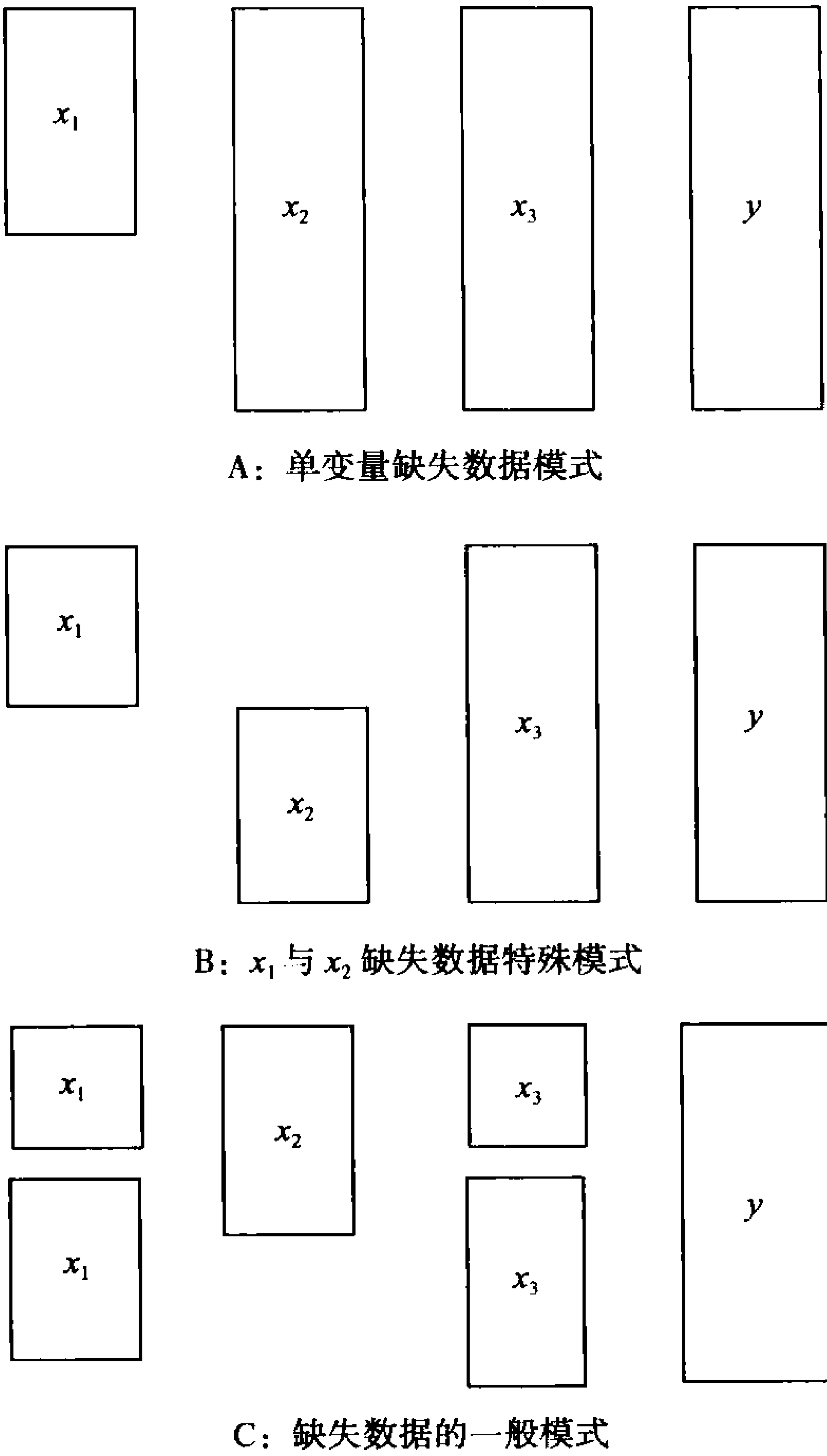


图 27.1 缺失数据：缺失回归元的例子

处理缺失数据的一种最简单方法是删失缺失数据，然后仅仅分析简化的"完整"观测值样本。例如，在 A 组调查对象中，完整样本是由 $\mathbf{x}_1$ 的所有可利用数据构成的 $(\mathbf{y}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ 的子集。不过，在 B 组调查对象中，当沿用这种方法时，人们舍弃无用的观测值，除非人们从分析中去掉 $(\mathbf{x}_1, \mathbf{x}_2)$。在 C 组调查对象中，完整数据集是在删除任何包括三个回归元当中任一个出现缺失数据的观测值之后形成的。

上述方法称为成列删除(**listwise deletion**)。该方法已被广泛采用，并且经常是统计软件的默认选项。此方法不一定无害处；其结果依赖于缺失数据机制，而且从这种研究中得出的结论可能出现严重缺陷。当然，一般地讲，丢掉数据意味着丢掉信息，同时将降低估计效率。因此，倘若归因于缺失数据的缺口以能不产生曲解的

方式得以填补，则成列删除看起来似乎值得尝试，本章将研究其他一些可供选择的方法及其局限性。

广义地讲，估算存在两种方法，一种是基于模型方法（**model-based**），另一种则不是基于模型方法。第一种方法运用模型对缺失观测值加以设定，然后使用得到的完整数据集去获得模型参数的更好估计。该过程是反反复复的。单个估算与多重估算都是可行的。现代方法的一个重要特征是，将缺失数据处理成随机变量，然后用从假定的基本分布中抽取的多重值来代替，这种过程称为多重估算（**multiple imputation**）。利用模拟方法可逼近这类分布。

由于估算是微观经济计量研究中的一个重要方面，所以有必要将此专题作为一个独立又简短的介绍章节。调查数据不可避免地包含缺失数据，一种普遍做法即成列删除就是一个估算方法。还可利用更好的估算方法。不过，我们应该注意到：所有估算方法均建立在假设基础上，而在某些应用中，有些估算假设是根本站不住脚的。

本章大部分内容研究基于模型方法。27.2 节介绍估算文献里占据主导地位的术语与假设。27.3 节给出不使用模型来处理缺失数据方法的一个简述。27.4 节首先从基于模型的方法开始，然后讨论极大似然法。27.5 节考察估算的回归框架及 EM 形式方法。27.6～27.7 节阐述利用数据增广的贝叶斯方法与 MCMC 进行估算的方法。27.8 节给出一个说明性例子。27.6～27.8 节提出运用第 13 章贝叶斯方法的一个精彩应用。

27.2 缺失数据假设

估算文献广泛使用的某些基本术语与正式定义都要归功于鲁宾（Rubin，1976）的研究工作，他曾引入两种重要的缺失数据机制，一种是随机缺失，另一种是完全随机缺失，这两种机制成为有用的基准。

鲁宾的设置背景包括 $\mathbf{Y}$，$\mathbf{Y}$ 是由完整数据集构成的 $N\times p$ 阶矩阵，它可能不是全部被观测到的。用 $\mathbf{Y}_{\text{obs}}$ 表示观测部分，$\mathbf{Y}_{\text{mis}}$ 表示不可观测到（缺失）部分。在回归模型背景下，$\mathbf{Y}$ 既可以指回归元，又可以指响应（因）变量。因此，该分析涵盖了缺失数据的一般情况。设 $\mathbf{R}$ 表示指示变量的 $N\times p$ 阶矩阵，$\mathbf{R}$ 的元素是 0 或 1，这要依据 $\mathbf{Y}$ 中的对应值是缺失的还是观测的。

对于含有单个因变量的回归来说，$\mathbf{Y}$ 包括响应变量 $\mathbf{y}$ 与 $(p-1)$ 个回归元 $\mathbf{X}$ 的数据。变量 x_k 的第 i 个观测值记为 x_{ki}，缺失的概率可能是下述情形：(i) 与其实际值独立；(ii) 依赖于其实际值；(iii) 依赖于 x_{kj}，$j\neq i$；(iv) 依赖于 x_{lj}，$j\neq i, l\neq k$。

下面给出关于缺失结构的假设。

27.2.1 *随机缺失*

设 $x_i(i=1,\cdots,N)$ 表示所研究数据集中的变量观测值。随机缺失假设[**missing at random (MAR) assumption**]是指如下的缺失情况，x_i 缺失并不依赖于 x_i 值，但可能依赖于 $x_j(j\neq i)$ 值。正式地讲：

$$x_i \text{ 是 MAR 的} \Rightarrow \Pr[x_i \text{ 出现缺失} | x_i, x_j, \forall j \neq i] \quad (27.1)$$
$$\Rightarrow \Pr[x_i \text{ 出现缺失} | x_j, \forall j \neq i]$$

在控制 x 的其他观测值之后，x_i 出现缺失的概率与 x_i 之值不相关。

鲁宾(Rubin，1976)给出的更为正式的定义可表述如下：MAR 假设蕴含着指示变量 **R** 的概率模型并不依赖于 $\mathbf{Y}_{mis}$，即：

$$\Pr[\mathbf{R} | \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \psi] = \Pr[\mathbf{R} | \mathbf{Y}_{obs}, \psi]$$

其中，ψ 表示缺失机制的基本(向量) 参数。

在 MAR 条件下，无响应包括在忽略缺失信息机制的基于似然推断之中，尽管所得到的估计值可能是无效的。可是，若 MAR 假设失效，则缺失概率依赖于不可观测的缺失值。由于缺失数据的值是未知的，所以 MAR 约束不是可检验的。因为缺失数据值是未知的。由于 MAR 是一个强假设，所以基于缺失性各种不同假设的敏感性分析是值得做的。

一个单独问题是，缺失数据模式是否是纯随机的。在实际应用中，我们希望观测值缺失处于数据聚集内部，在第 24 章的意义下，观测值可能是相关的。可是，该问题并不与因缺失性及数据值有联系而产生的无响应偏倚有关。

27.2.2 *完全随机缺失*

完全随机缺失(**Missing completely at random**，记为 **MCAR**)是 MAR 的一种特殊情况。它意味着，$\mathbf{Y}_{obs}$ 是所有潜在可观测数据值的一个简单随机样本[谢弗(Schafer，1997)]。

再次假定 x_i 是正在研究的数据集中变量的一个观测值。于是，x_i 的数据被称为 MCAR，如果 x_i 缺失数据的概率既不依赖于 x_i 之值，也不依赖于数据中其他变量的值。正式地讲：

$$x_i \text{ 是 MCAR 的} \Rightarrow \Pr[x_i \text{ 出现缺失} | x_i, x_j, \forall j \neq i] \quad (27.2)$$
$$\Rightarrow \Pr[x_i \text{ 出现缺失}]$$

例如，如果：(a) 平均地讲，没有报告收入的那些人比报告收入的人要年轻；(b) 典型小的(大的)值出现缺失，就违背了 MCAR。

对于本节前面所提及的情况(i)～(v)，情况(i)既满足 MCAR 又满足 MAR，情况(iii)与(iv)均满足 MAR，而情况(ii)则不满足 MAR。

MCAR 蕴含，观测数据是所有样本的一个随机子样本。当假设有效，因而忽略不完整观测值即观测值含有缺失值时，就不会产生偏倚。

一个推论是，MCAR 失效蕴含样本有选择偏倚形式。MAR 虽是一个较弱假设，但仍有助于估算，这是因为它假定缺失数据机制仅依赖于观测量。

27.2.3 *可忽略缺失与不可忽略缺失*

缺失数据机制被称为可忽略的(**ignorable**)，如果：(a) 数据集是 MAR 的；(b) 缺失数据生成过程的参数 ψ 与我们要估计的参数不相关。

这个条件类似于第 2 章曾讨论的弱外生性条件，意味着模型参数 $\boldsymbol{\theta}$ 与缺失机制参数 ψ 截然不同。因而，如果缺失数据是可忽略的，就不需要将缺失数据的缺口建模成建立模型演算的一个基础性部分。在可忽略缺失条件 (b)几乎总是得到满足的假设下，MAR 与“可忽略性”经常被处理成等价的[阿利森(Allison, 2002)]。

如果对于(y,x)来说，MAR 假设被违背了，就产生非可忽略的缺失数据机制，但是若仅对 x 来说被违背时，则没有违背非可忽略的缺失数据机制。在这种情况下，为了获得参数 $\boldsymbol{\theta}$ 的一致估计，必须对缺失数据生成过程以及整个模型加以建模。为了避免选择偏倚的可能性，必须使用诸如赫克曼两阶段方法的估计量。

估算文献关注可忽略缺失性。若数据集是 MCAR，则撇开可通过估算减少的效率损失不谈，缺失数据并不会引起什么问题。相反，如果数据集仅仅是 MAR，那么为确保一致性与提高效率或许必用估算方法。

27.3 非模型处理缺失数据

倘若没有模型可以利用，则人们直接分析可用数据，或者分析非模型估算之后的数据。

27.3.1 *只利用可用数据*

成列删除或完整个案分析意指，删除数据中有缺失值的一个或多个变量的那种观测值(个案)。在 MCAR 假设下，经过成列删除之后，所保留的样本仍是源自最初总体的一个随机样本；因此，基于该样本的估计是一致的。不过，其标准误差将会扩大，因为所用信息甚少。若回归元个数很多，则成列删除的总效果导致总观测值会剧烈减少。这激发人们脱离那种对拥有高比例缺失观测值的变量进行分析，可是，由该种方法所产生的结果却潜在地对人误导。

如果 MCAR 得不到满足且缺失数据仅仅是 MAR，那么估计将是有偏的。因而，成列删除对违背 MCAR 而言不是稳健的。不过，成列删除对回归分析中各个自变量(回归元)违背 MAR 而言是稳健的，也就是说，任何回归元出现缺失数据的概率并不依赖于因变量之值。简略地讲，成列删除是可接受的，如果归因于缺失数据的不完全情况构成了各种情况的比例很小，比如说 5% 或更少[谢弗(Schafer, 1996)]。重要的是，成列删除之后的样本是所研究总体的代表。

成对删除(**pairwise deletion**)或可用案例分析，时常被认为是比成列删除更好的一种方法。其思想是估计(x_1, x_2)的联合样本矩时，运用观测值(x_{1i}, x_{2i})的全部可能对，并且估计边缘矩时运用个体变量的全部观测值。因而，在线性回归中，在成对删除下我们运用回归元的所有可能对估计 $(\mathbf{X}'\mathbf{X})$与$(\mathbf{X}'\mathbf{y})$，而在成列删除下，要在删除任何拥有缺失观测值的全部情况后才能估计$(\mathbf{X}'\mathbf{X})$与$(\mathbf{X}'\mathbf{y})$。很明显，在成对删除下，我们损失较少信息。这里建议要运用最大信息量去估计个体概括统计量，诸如均值与协方差，然后使用这些概括统计量去计算回归估计。

成对删除有两个重要局限性：(1) 一般地讲，估计标准误差与检验统计量都是有偏的；(2) 所得到的回归元协方差矩阵 $(\mathbf{X}'\mathbf{X})$ 可能不是正定的。

27.3.2 *不用模型的估算*

统计软件经常执行一系列专门或勉强证明合理的方法。

均值估算（**mean imputation**）或均值替补（**mean substitution**）意指，运用可利用值的平均值代替缺失观测值。该方法是均值保留，但将对数据的边缘分布产生影响。很明显，边缘分布中心概率质量表现出增大。该方法也影响到协方差以及与其他变量的相关性。

简单替补[1]（**simple hot deck**）估算意指，用从有观测值的变量中随机抽取到的值代替缺失值，这有点像自助法。该方法维持了那个变量的边缘分布，却扭曲了变量之间的协方差与相关性。

在回归背景下，这两个著名方法虽然具有简单性，但它们没有一个引人注目。

27.4 观测数据似然函数

缺失数据的现代方法是，通过从基于假定观测数据模型或缺失数据机制中抽取的单个或多重值来估算缺失观测值。这种方法的贝叶斯变形是从后验分布中采样，既使用似然函数又使用参数的先验分布。

第一个重要问题涉及估算方法中缺失数据机制所起的作用，特别是，缺失数据机制是否是可忽略的。

设 $\boldsymbol{\theta}$ 表示 $\mathbf{Y}=(\mathbf{Y}_{obs},\mathbf{Y}_{mis})$ 数据生成过程的参数，并设 $\boldsymbol{\psi}$ 表示缺失数据机制的参数。为了符号简单起见，假定 $(\mathbf{Y}_{obs},\mathbf{Y}_{mis})$ 均是连续变量。于是，$(\mathbf{R},\mathbf{Y}_{obs})$ 的联合分布由

$$
\begin{aligned}
\Pr[\mathbf{R},\mathbf{Y}_{obs}\mid\boldsymbol{\theta},\boldsymbol{\psi}] &= \int\Pr[\mathbf{R},\mathbf{Y}_{obs},\mathbf{Y}_{mis}\mid\boldsymbol{\theta},\boldsymbol{\psi}]\,\mathbf{dY}_{mis} \qquad (27.3)\\
&= \int\Pr[\mathbf{R}\mid\mathbf{Y}_{obs},\mathbf{Y}_{mis},\boldsymbol{\psi}]\Pr[\mathbf{Y}_{obs},\mathbf{Y}_{mis}\mid\boldsymbol{\theta}]\mathbf{dY}_{mis}\\
&= \Pr[\mathbf{R}\mid\mathbf{Y}_{obs},\boldsymbol{\psi}]\int\Pr[\mathbf{Y}_{obs},\mathbf{Y}_{mis}\mid\boldsymbol{\theta}]\,\mathbf{dY}_{mis}\\
&= \Pr[\mathbf{R}\mid\mathbf{Y}_{obs},\boldsymbol{\psi}]\Pr[\mathbf{Y}_{obs}\mid\boldsymbol{\theta}]
\end{aligned}
$$

给出，其中，第一个等式是从所有数据与 $\mathbf{R}$ 的联合概率中通过对 $\mathbf{Y}_{mis}$ 进行积分（或者平均），进而推导出 $(\mathbf{R},\mathbf{Y}_{obs})$ 的联合概率。第二行将联合概率因式分解为以 $\mathbf{Y}_{obs}$ 与 $\mathbf{Y}_{mis}$ 为条件的条件成分及边缘成分。第三行从观测数据机制中分离出缺失数据机制；该步由 MAR 假设得出。最后一行意味着，$\boldsymbol{\theta}$ 与 $\boldsymbol{\psi}$ 是截然不同的参数，从而对 $\boldsymbol{\theta}$ 进行推断能忽略缺失数据机制，而仅仅依赖于 $\mathbf{Y}_{obs}$。

观测数据似然是与第四行的最后因子成比例：

$$
\mathrm{L}[\boldsymbol{\theta}\mid\mathbf{Y}_{obs}]\propto\Pr[\mathbf{Y}_{obs}\mid\boldsymbol{\theta}] \qquad (27.4)
$$

该观测数据似然只涉及观测数据 $\mathbf{Y}_{obs}$，尽管参数 $\boldsymbol{\theta}$ 出现在全部观测值（观测到数据与缺失数据）的数据生成过程中。正如第 13 章一样，比例常值没有出现在式

[1] 又称为热平台法。——译者注

(27.4)之中。

在 MAR 假设下，$(\boldsymbol{\theta},\boldsymbol{\psi})$的联合后验概率可被写成 $\Pr[\mathbf{R},\mathbf{Y}_{obs}|\boldsymbol{\theta},\boldsymbol{\psi}]$与先验联合分布 $\pi(\boldsymbol{\theta},\boldsymbol{\psi})$的如下乘积形式：

$$\begin{aligned}\Pr[\boldsymbol{\theta},\boldsymbol{\psi}\,|\,\mathbf{Y}_{obs},\mathbf{R}] &= k\Pr[\mathbf{R},\mathbf{Y}_{obs}\,|\,\boldsymbol{\theta},\boldsymbol{\psi}]\pi(\boldsymbol{\theta},\boldsymbol{\psi}) \\ &\propto \Pr[\mathbf{R}\,|\,\mathbf{Y}_{obs},\boldsymbol{\psi}]\Pr[\mathbf{Y}_{obs}\,|\,\boldsymbol{\theta}]\pi(\boldsymbol{\theta},\boldsymbol{\psi}) \\ &\propto \Pr[\mathbf{R}\,|\,\mathbf{Y}_{obs},\boldsymbol{\psi}]\Pr[\mathbf{Y}_{obs}\,|\,\boldsymbol{\theta}]\pi_\theta(\boldsymbol{\theta})\,\pi_\psi(\boldsymbol{\psi})\end{aligned} \tag{27.5}$$

其中，第一行中的 k 表示与$(\boldsymbol{\theta},\boldsymbol{\psi})$无关的一个比例性常值。第二行用到了式(27.3)给出的因式分解，而第三行则使用了 $\boldsymbol{\theta}$ 与 $\boldsymbol{\psi}$ 是独立先验的假设。

因为主要关注内容在于 $\boldsymbol{\theta}$，所以从联合后验中通过对 $\boldsymbol{\psi}$ 进行积分，推导 $\boldsymbol{\theta}$ 的边缘后验。从而得出观测数据后验(**observed-data posterior**)：

$$\begin{aligned}\Pr[\boldsymbol{\theta}\,|\,\mathbf{Y}_{obs},\mathbf{R}] &= \int \Pr[\boldsymbol{\theta},\boldsymbol{\psi}\,|\,\mathbf{Y}_{obs},\mathbf{R}]d\boldsymbol{\psi} \\ &\propto \Pr[\mathbf{Y}_{obs}\,|\,\boldsymbol{\theta}]\pi_\theta(\boldsymbol{\theta})\int \Pr[\mathbf{R}\,|\,\mathbf{Y}_{obs},\boldsymbol{\psi}]\pi_\psi(\boldsymbol{\psi})d\boldsymbol{\psi} \\ &\propto \mathrm{L}[\boldsymbol{\theta}\,|\,\mathbf{Y}_{obs}]\pi_\theta(\boldsymbol{\theta})\end{aligned} \tag{27.6}$$

其中，第二行将 $\boldsymbol{\theta}$ 与 $\boldsymbol{\psi}$ 分离开，而最后一行将积分表达式合并到比例性常值之中。因此，最后一行没有包含 $\boldsymbol{\psi}$，从而与缺失数据机制 $\mathbf{R}$ 独立。

27.5 基于回归的估算

在本节，我们考察基于最小二乘法的估算。其重要组成部分是运用 EM 算法，前面曾引进 EM 算法，并在 10.3.7 节讨论过。

EM 算法由期望步骤与求极大值步骤组成。EM 算法的结构与贝叶斯 MCMC 以及数据扩大方法紧密地联系。因此，我们将引入一个例子，阐述支撑现代多重估算方法的动因，并给出这类方法的重要特性，而不是提供处理缺失数据的完整操作方法。

27.5.1 *因变量出现缺失数据的线性回归例子*

在实际应用中，因变量(内生变量)与/或者解释变量可能出现缺失观测值。我们考察一个回归例子，其中因变量有缺失观测值，即：

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_{mis}\end{bmatrix}=\begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{bmatrix}\boldsymbol{\beta}+\begin{bmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{bmatrix} \tag{27.7}$$

其中，$\mathrm{E}[\mathbf{u}\,|\,\mathbf{X}]=\mathbf{0}$，$\mathrm{E}[\mathbf{u}\mathbf{u}'\,|\,\mathbf{X}]=\sigma^2\mathbf{I}_N$。新出现的困难是，因变量 $\mathbf{y}$ 观测值的一部分出现缺失，将此缺失部分记为 $\mathbf{y}_{mis}$。我们假定，可利用的完全观测值是来自总体的一个随机样本，因而假定缺失数据虽不是 MCAR 的，却是 MAR 的。

已知 MAR 假设，且 $N_1>K$，N_1 的第一分块能用于一致地估计出 K 维参数。在高斯误差条件下，$(\boldsymbol{\beta},\sigma^2)$的极大似然估计是 $\hat{\boldsymbol{\beta}}=[\mathbf{X}_1'\mathbf{X}_1]^{-1}\mathbf{X}_1'\mathbf{y}_1$ 与 $s^2=(\mathbf{y}_1-\mathbf{X}_1\hat{\boldsymbol{\beta}})'(\mathbf{y}_1-\mathbf{X}_1\hat{\boldsymbol{\beta}})/N_1$。借助于标准理论知识，并在正态性假设下，得到 $\hat{\boldsymbol{\beta}}|$数据～

$\mathcal{N}[\boldsymbol{\beta}, \sigma^2[\mathbf{X}_1'\mathbf{X}_1]^{-1}]$与 $s^2/\sigma^2 \mid \hat{\boldsymbol{\beta}} \sim (N_1-K)\chi^2_{N_1-K}$。

首先，考察生成缺失观测值的一种朴素单一估算方法。以 $\mathbf{X}_2$ 为条件，$\mathbf{y}_{\text{mis}}$的预测值记为 $\hat{\mathbf{y}}_{\text{mis}}$，$\hat{\mathbf{y}}_{\text{mis}}$由 $\mathbf{X}_2\hat{\boldsymbol{\beta}}$ 给出，其中，$\hat{\boldsymbol{\beta}}$ 表示仅利用前面 N_1 个观测所获得的先前估计值。于是：

$$\hat{\mathrm{E}}[\mathbf{y}_{\text{mis}} \mid \mathbf{X}_2] = \hat{\mathbf{y}}_{\text{mis}} = \mathbf{X}_2\hat{\boldsymbol{\beta}} \tag{27.8}$$
$$\hat{\mathrm{V}}[\hat{\mathbf{y}}_{\text{mis}}] \equiv \hat{\mathbf{V}}[\hat{\mathbf{y}} \mid \mathbf{X}_2] = s^2(\mathbf{I}_{N_2} + \mathbf{X}_2[\mathbf{X}_1'\mathbf{X}_1]^{-1}\mathbf{X}_2')$$

其中，$s^2\mathbf{I}_{N_2}$ 表示 $\mathbf{V}_{[\mathbf{u}_2]}$的估计值。

就上述简便方法而言，人们可生成 N_2 个 $\mathbf{y}_{\text{mis}}$的预测值，然后将标准回归方法应用到 $N=N_1+N_2$ 观测值的全部样本上。

简便方法的两个步骤对应于 EM 算法的两个步骤。预测步骤是 E 步骤，而将最小二乘应用于扩大样本的第二步则是 M 步骤。

不过，这种解显得不精细。第一，考虑数据扩大步骤。由于生成值 $\hat{\mathbf{y}}_{\text{mis}}$准确地位于最小二乘拟合平面上，为了得到一个新的估计值 $\hat{\boldsymbol{\beta}}_A$，将$(\hat{\mathbf{y}}_{\text{mis}}, \mathbf{X}_2)$加入样本之中并不会改变先前估计值 $\hat{\boldsymbol{\beta}}$：

$$\begin{aligned}\hat{\boldsymbol{\beta}}_A &= [\mathbf{X}_1'\mathbf{X}_1 + \mathbf{X}_2'\mathbf{X}_2]^{-1}[\mathbf{X}_1'\mathbf{y}_1 + \mathbf{X}_2'\hat{\mathbf{y}}_{\text{mis}}] \\ &= [\mathbf{X}_1'\mathbf{X}_1 + \mathbf{X}_2'\mathbf{X}_2]^{-1}[\mathbf{X}_1'\mathbf{X}_1\hat{\boldsymbol{\beta}} + \mathbf{X}_2'\mathbf{X}_2\hat{\boldsymbol{\beta}}] \\ &= \hat{\boldsymbol{\beta}}\end{aligned}$$

第二，因为由构造知，添加的 N_2 个残差均为 0，通过标准公式获得来自扩大样本的 σ^2 估计值，该估计值显得太小，即：

$$\begin{aligned}s_A^2 &= (\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_A)'(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_A)/N \\ &= (\mathbf{y}-\mathbf{X}_1\hat{\boldsymbol{\beta}})'(\mathbf{y}_1-\mathbf{X}_1\hat{\boldsymbol{\beta}})/N < s^2\end{aligned} \tag{27.9}$$

正确地讲，其中，s^2 应被 N_1 而不是 N 除。

最后，正如从 $\hat{\mathbf{y}}_{\text{mis}}$的抽样方差中看到的，与 $\mathbf{y}_1$ 不同，生成预测都是异方差的，从而 $\hat{\boldsymbol{\beta}}_A$ 的方差不能利用通常最小二乘公式加以估计。观测值 $\hat{\mathbf{y}}_{\text{mis}}$是从具有不同方差的分布中抽取的。这种简便方法没有考虑到依附于 $\hat{\mathbf{y}}_{\text{mis}}$估计的不确定性。

为了确定这些问题，就需要校正。首先，$\hat{\mathbf{y}}_{\text{mis}}$的估计应考虑到 $\hat{\boldsymbol{\beta}}$ 的不确定性。通过调整 $\hat{\mathbf{y}}_{\text{mis}}$可达到此目的，并将某些“噪声”加入生成预测之中，使得缺失数据估计值更紧密地酷似从 $\mathbf{y}_1$ 的(估计或条件)分布所抽取的值。标准化步骤用到了下面事实：$\mathrm{V}[\hat{\mathbf{y}}_{\text{mis}}]$的估计值 $\hat{\mathbf{V}}$ 可从式(27.8)得出。因此，变换变量 $\hat{\mathbf{V}}^{-1/2}\hat{\mathbf{y}}_{\text{mis}}$的成分拥有单位方差。为了类似 $\mathbf{y}_1$ 的分布，我们运用从 $\mathcal{N}[0, s^2]$分布实施蒙特卡罗抽样，并用 $\hat{\mathbf{V}}^{-1/2}\hat{\mathbf{y}}_{\text{mis}}$乘以它。

修正算法如下：

1. 利用前面 N_1 个完整观测值估计值 $\hat{\boldsymbol{\beta}}$。

2. 生成 $\hat{\mathbf{y}}_{\text{mis}} = \mathbf{X}_2\hat{\boldsymbol{\beta}}_2$。

3. 生成 $\hat{\mathbf{y}}^a_{\text{mis}} = (\hat{\mathbf{V}}^{-1/2}\hat{\mathbf{y}}_{\text{mis}}) \odot \mathbf{u}_m$ 的调整值，其中，$\mathbf{u}_m$ 表示由 $\mathcal{N}[0, s^2]$分布得出的蒙特卡罗抽样值，而$\odot$表示元素对元素逐一乘法。

4. 运用扩大样本得到$\hat{\beta}$的修正估计值。

5. 重复步骤 1～4，步骤 1 将用到$\hat{\beta}$的修正估计。

修正算法也称为 EM 类型算法，该方法不断实施，一直到它在下述情况收敛为止：系数变化或回归残差平方和变化可任意小。

为了与刚才讨论的内容连接上，我们对该算法给出一种不同解释。第 3 步是从给定 β 时 y 的条件分布中抽取，而第 4 步是从给定 s^2、$\mathbf{X}$ 时从 β 的条件分布中抽取的。这种方法可通过增加如下一步而得到精炼，即增加一步是从 s^2 分布中抽样。我们没有做完该方法的所有步骤，因为在后面对估算的讨论中会变更为清楚。

第 16 章曾阐述过因变量出现缺失数据时的另一些模型。这些模型放松了 MAR 假设，并设定非忽略缺失性。于是，用上述 EM 算法，就得到 β 的非一致估计。删失 Tobit 模型设定：对于满足 $\mathbf{x}'\beta+u\leqslant 0$ 观测值来说，数据出现缺失，而且一个一致估计量是 Tobit 极大似然估计量（参见 16.3 节）。雨宫（Amemiya，1985，第 376～378 页）曾经详述过 Tobit 模型的 EM 算法。

27.6 数据扩大与 MCMC

缺失数据贝叶斯方法的一般性结构运用了下述形式的迭代算法，即用估算步骤与预测步骤。

估算步骤（**imputation step，I** 步）是从 $\mathbf{Y}_{\text{mis}}$ 的条件预测分布抽样。已知第 r 回的估计值：

$$\mathbf{Y}_{\text{mis}}^{(r+1)}\sim \Pr[\mathbf{Y}_{\text{mis}}\mid \mathbf{Y}_{\text{obs}},\boldsymbol{\theta}^{(r)}] \tag{27.10}$$

这个表达式给定当前估计值 $\boldsymbol{\theta}^{(r)}$ 与观测数据 $\mathbf{Y}_{\text{obs}}$ 时，从 $\mathbf{Y}_{\text{mis}}$ 的预测条件分布随机抽样得到 $\mathbf{Y}_{\text{mis}}^{(r+1)}$。注意，一般地讲，$\mathbf{Y}_{\text{mis}}$ 是一个矩阵，故这样符号（原则上）涉及到一系列抽样。

预测步骤（**prediction step，P** 步）是通过从完整数据后验

$$\boldsymbol{\theta}^{(r+1)}\sim \Pr[\boldsymbol{\theta}\mid \mathbf{Y}_{\text{obs}},\mathbf{Y}_{\text{mis}}^{(r+1)}] \tag{27.11}$$

抽样而完成的。也就是说，$\mathbf{Y}_{\text{obs}}$ 借助于从 $\mathbf{Y}_{\text{mis}}$ 预测分布抽样得到估算值 $\mathbf{Y}_{\text{mis}}^{(r+1)}$ 得到扩大，然后，从 $\boldsymbol{\theta}$ 的后验分布得到一个抽样。对式（27.10）与式（27.11）步骤不断重复进行。

从两个分布中得到的抽样序列生成了马尔可夫链。这样过程非常类似 EM 算法，本质上是 13.5.2 节的吉布斯抽样器，可是在缺失数据文献中，它称为数据扩大（**data augmentation**）。在适当条件下，并借助于 13.5.1 节所引述的定理，对于充分大的 r 值来说，抽样序列将收敛到平稳分布，r 为此链长度。当该链终止时，我们就有 $\mathbf{Y}_{\text{mis}}$ 的一个估算。于是，将 $\boldsymbol{\theta}^{(r)}$ 看成是从 $\Pr[\boldsymbol{\theta}\mid \mathbf{Y}_{\text{obs}}]$ 中抽样得出的一个近似，而 $\mathbf{Y}_{\text{mis}}^{(r+1)}$ 是从 $\Pr[\mathbf{Y}_{\text{mis}}\mid \mathbf{Y}_{\text{obs}}]$ 中抽样得到的一个近似。如同任何 MCMC 应用一样，该链为确保后继估算没有统计相依性而必须执行得充分长。这些问题已在第 13 章讨论过。

在收敛之后，我们可以完成如下两个联合目标：一个是基于数据的设定模型估算缺失值，另一个是利用观测值与估算估计模型。一旦收敛，我们就拥有必须计算

$\boldsymbol{\theta}$ 的后验矩以及 $\boldsymbol{\theta}$ 与 $\mathbf{Y}$ 的任何关注函数的数据，其所用思想已在第 13 章讨论了。

作为这个方法的一个解释，我们重新考察前一节缺失数据回归的例子。MCMC 算法的步骤如下：

1. 利用观测数据，计算 $\hat{\boldsymbol{\beta}}=[\mathbf{X}_1'\mathbf{X}_1]^{-1}\mathbf{X}_1'\mathbf{y}_1$，以及 $\hat{\mathbf{u}}=(\mathbf{y}_1-\mathbf{X}_1\hat{\boldsymbol{\beta}})$。
2. 当用从 $\chi^2_{N_1-K}$ 分布中得到的抽样除 $\hat{\mathbf{u}}'\hat{\mathbf{u}}$ 时，就生成了 σ^2。
3. 从 $\boldsymbol{\beta}|\sigma^2\sim\mathcal{N}[\hat{\boldsymbol{\beta}},\sigma^2[\mathbf{X}_1'\mathbf{X}_1]^{-1}]$ 中得到抽样。
4. 从 $\mathbf{Y}_{\mathrm{mis}}\sim\mathcal{N}[\mathbf{X}_2\hat{\boldsymbol{\beta}},\sigma^2]$[1]中得到抽样。
5. 用 $\mathbf{y}$ 代替 $\mathbf{y}_1$，用 $\mathbf{X}$ 代替 $\mathbf{X}_1$，在进行适当调整后，重复步骤 1～4。

执行步骤 2 的理由是，在($\boldsymbol{\beta},\sigma^2$)的非信息先验条件下，只有使用观测数据时，$\hat{\mathbf{u}}'\hat{\mathbf{u}}/\sigma^2$ 的条件后验分布服从 $\chi^2_{N_1-K}$。在数据扩大之后，这就变成 χ^2_{N-K}。执行步骤 3 的理由是，在非信息先验条件下，条件后验分布服从 $\mathcal{N}[\hat{\boldsymbol{\beta}},\sigma^2[\mathbf{X}_1'\mathbf{X}_1]^{-1}]$。一旦数据扩大后，这变成 $\mathcal{N}[\hat{\boldsymbol{\beta}},\sigma^2[\mathbf{X}'\mathbf{X}]^{-1}]$。步骤 4 则是使用条件预测密度 $\mathcal{N}[\mathbf{X}_2\hat{\boldsymbol{\beta}},\sigma^2]$ 的估算步骤。倘若我们使用信息先验，例如($\boldsymbol{\beta},\sigma^2$)的正态伽玛先验，则这些步骤就要进行适当修正。这种情况的条件后验分布已由 13.3 节给出。

27.7 多重估算

前面一节分析了如何实施完整的 MCMC 生成单一估算。不过，单一估算并不适合处理缺失数据的不确定性。这就是要使用多重估算方法的根本原因。$\mathbf{Y}_{\mathrm{mis}}|\mathbf{Y}_{\mathrm{obs}},\boldsymbol{\theta}$ 的条件预测分布可通过对 $\boldsymbol{\theta}$ 的观测数据后验加以平均而获得：

$$\Pr[\mathbf{Y}_{mis}|\mathbf{Y}_{obs}]=\int\Pr[\mathbf{Y}_{mis}|\mathbf{Y}_{obs},\boldsymbol{\theta}]\Pr[\boldsymbol{\theta}|\mathbf{Y}_{obs}]\mathbf{d}\boldsymbol{\theta}$$

给定模型参数的不确定性，从贝叶斯观点来看，适当的多重估算反映出 $\mathbf{Y}_{\mathrm{mis}}$ 的不确定性。

在多重估算后，缺失数据 $\mathbf{Y}_{\mathrm{mis}}$ 就用模拟/估算值 $\mathbf{Y}_{\mathrm{mis}}^{(1)},\mathbf{Y}_{\mathrm{mis}}^{(2)},\mathbf{Y}_{\mathrm{mis}}^{(3)},\cdots,\mathbf{Y}_{\mathrm{mis}}^{(m)}$ 代替。那么，每一个完整数据集都要得到分析，就好像数据集是完全的。经过 m 次分析得出结果，将显示由缺失数据引起的不确定性方面的变化。就 m 个不同数据集而言，会产生下述问题：人们应该如何确定一个合适的 m 值，同时应该如何将参数估计的 m 个集合与协方差矩阵结合起来。我们对这两个问题都要给予讨论，运用来自文献的一些结果，却没有提供所用结果的详细推导。

在考虑如何对基于多重估算数据的一些结果加以结合方面，一个重要结果可用任意统计量 Q 来进行表述，即：

$$\Pr[Q|\mathbf{Y}_{\mathrm{obs}}]=\int\Pr[Q|\mathbf{Y}_{\mathrm{mis}},\mathbf{Y}_{\mathrm{obs}}]\Pr[\mathbf{Y}_{\mathrm{mis}}|\mathbf{Y}_{\mathrm{obs}}]\mathbf{d}\mathbf{Y}_{\mathrm{mis}}\tag{27.12}$$

该式描述了 Q 的实际后验分布，式(27.12)通过对 Q 的完整数据后验分布进行平均而得到。这意味着在缺失观测值的多重估算结果上取平均值。

[1] 原著中此处为([$\mathbf{X}_2\hat{\boldsymbol{\beta}}$, σ^2]，但应该为([$\mathbf{X}_2\hat{\boldsymbol{\beta}},\sigma^2\mathbf{I}$]。——译者注

式(27.12)蕴含,Q 的最终估计量通过期望迭代律给出:

$$E[Q|\mathbf{Y}_{obs}]=E[E[Q|\mathbf{Y}_{obs},\mathbf{Y}_{mis}]|\mathbf{Y}_{obs}] \tag{27.13}$$

Q 的后验均值是利用缺失数据的重复估算后的完整数据而得到 Q_r 的平均值。

Q 的最终方差由公式

$$V[Q|\mathbf{Y}_{obs}]=E[V[Q|\mathbf{Y}_{obs},\mathbf{Y}_{mis}]|\mathbf{Y}_{obs}]+V[E[Q|\mathbf{Y}_{obs},\mathbf{Y}_{mis}]|\mathbf{Y}_{obs}] \tag{27.14}$$

给出,这里用到了 A.8 节给出的方差分解公式。

鲁宾(Rubin, 1996)还给出了整合矩信息的下述规则,这可用纯量参数加以表述。对于任意纯量参数,假定 $\hat{Q}_r$ 是第 r 回估算的点估计,$\hat{U}_r$ 是方差估计。于是,可分别定义出如下点估计与方差估计的平均值:

$$\bar{Q}=m^{-1}\sum_{r=1}^{m}\hat{Q}_r \tag{27.15}$$

$$\bar{U}=m^{-1}\sum_{r=1}^{m}\hat{U}_r \tag{27.16}$$

而将估算之间的方差定义成:

$$B=(m-1)^{-1}\sum_{r=1}^{m}(\hat{Q}_r-\bar{Q})^2 \tag{27.17}$$

并且总方差定义成:

$$T=\bar{U}+(1+m^{-1})B \tag{27.18}$$

结果(27.15)与式(27.16)可由式(27.13)得出;而式(27.18)则由式(27.14)得到。谢夫(Schafer, 1997)给出将 p 值与似然比统计量结合起来的结果,并提供另外一些参考文献。

利用最终估计,可做出关于个体系数或系数子集的估算后推断,这是因为标准的中心极限定理与有关的大样本结论均能推广到涵盖此情况。

下面是关于 m 重估算的相对效率的一个测量:

$$reff=(1+(\lambda/m))^{-1} \tag{27.19}$$

其中,λ 是缺失观测值的比例小数。测量是相对于没有缺失数据而给出的。表 27.1 的算术计算结果表明,仅就三个估算而言,对于缺失数据有 10%时,其效率高达 97%,而对于缺失数据有 50%时,其效率为 86%。就 10 次或更多次估算而言,对于缺失数据有 50%时,其相对效率大于 95%。因而,正如谢弗(Schafer, 1997)强调的,估算次数不必太大。

表 27.1 多重估算的相对效率

	观测值缺失(λ)		
估算次数(m)	10%	30%	50%
3	0.967	0.909	0.857
10	0.990	0.970	0.952
20	0.995	0.985	0.975

27.8 缺失数据的估算例子

本节对缺失数据估算的两个应用给出解释:一个是成列删除与均值估算的无模型方法(参见 27.2 节),另一个是利用 MCMC 算法数据增广的基于模型方法(参见 27.6 节)。仅有回归元出现缺失,而缺失机制是 MAR 的。

第一个应用涉及简单多重估算,第二个应用涉及 logit 回归。为了清晰简单起见,我们使用已知数据生成过程来人为地生成数据。

27.8.1 *回归元出现数据缺失的线性回归*

对于线性回归例子,数据生成过程是:

$$y_i=\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+u_i, \quad i=1,2,\cdots,N \tag{27.20}$$

其中,$u_i|x_{1i},x_{2i}\sim\mathcal{N}[0,\sigma^2]$,$(x_{1i},x_{2i})$服从二元正态分布,满足:

$$\begin{bmatrix}x_{1i}\\x_{2i}\end{bmatrix}\sim\mathcal{N}\left[\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1&\rho\\\rho&1\end{bmatrix}\right] \tag{27.21}$$

所以 $x_{2i}|x_{1i}\sim\mathcal{N}[\rho x_{1i},1-\rho^2]$。而且,我们设 $\boldsymbol{\beta}'=[1\ \ 1\ \ 1]$, $N=1\,000$,x_1 与 x_2 出现随机缺失数据的比例为 10%或 25%。对于任意 i 来说,或 x_1 出现缺失或 x_2 出现缺失,或两者都出现缺失。我们还用到 ρ 的两个不同值,即 0.36 与 0.64。

就马尔可夫链而言,我们对“练习阶段”运用 500 次迭代。利用 SAS MI Proc 算法完成该马尔可夫链的计算,执行算法时用到了非信息先验。对于只是示范性目的来说,估算次数固定在 10 次,但该链在“练习阶段”之后的长度变动从 10 到 10 000。MI Proc 使用式(27.15)~(27.18)对来自多重估算的结果加以联合。

表 27.2 与表 27.3 阐述了 ρ 很大时,缺失数据出现大比例与小比例不同情况的结果。这些不同结果之间并没有巨大差异。因为应用到 MAR 假设,所以来自成列删除的点估计值与来自完整样本的点估计值接近,只是如人们所料,在成列删除条件下,标准误差较大。在均值估算条件下,β_2 的点估计相对而言更为发散,可是观测变异仍位于抽样误差界之内。很明显,在这两种情况下,马尔可夫链更迅速地达到平稳,在迭代 10 次与迭代 10 000 次之间,其结果差异很小。这可能归因于拥有练习阶段次数 500 的集合,比相对简单情况所需的更大一些。

表 27.2 缺失数据估算:出现 10%比例缺失数据与高相关性的线性估计,运用 MCMC 算法

				马尔可夫链的长度			
	无数据缺失	成列删除	均值估算	10	1 000	5 000	10 000
$\hat{\beta}_0$	0.919	0.913	0.899	0.910	0.911	0.909	0.903
	(0.104)	(0.113)	(0.105)	(0.102)	(0.101)	(0.103)	(0.101)
$\hat{\beta}_1$	1.097	1.067	1.053	1.196	1.205	1.199	1.199
	(0.138)	(0.167)	(0.150)	(0.148)	(0.155)	(0.144)	(0.147)
$\hat{\beta}_2$	1.000	1.072	1.112	1.042	1.051	1.041	1.055
	(0.132)	(0.145)	(0.135)	(0.140)	(0.146)	(0.143)	(0.146)
R^2	0.240	0.254	0.226				

表 27.3 缺失数据估算:对出现 25%的缺失数据与高相关性的线性回归估计,运用 MCMC 算法

	无数据缺失	成列删除	均值估算	马尔可夫链的长度			
				10	1 000	5 000	10 000
$\hat{\beta}_0$	0.919	0.863	0.984	0.899	0.898	0.925	0.900
	(0.104)	(0.167)	(0.108)	(0.108)	(0.105)	(0.111)	(0.110)
$\hat{\beta}_1$	1.097	1.048	1.062	1.028	1.047	1.082	0.987
	(0.138)	(0.167)	(0.150)	(0.152)	(0.166)	(0.161)	(0.155)
$\hat{\beta}_2$	1.000	1.129	1.156	1.071	1.085	1.024	1.124
	(0.132)	(0.161)	(0.148)	(0.152)	(0.144)	(0.172)	(0.152)
R^2	0.240	0.268	0.203				

表 27.4 表明,模拟练习重现了关于小 ρ 值且缺失数据为 25%的"最坏情况"。来自完整样本的点估计值与来自成列删除及均值估计情况的那些点估计之间的差异,从总体上看相对大于 MCMC 情况。不过,甚至在此情况下,由完整样本得出的估计值之间,实际上并不存在引人注目的差别。我们再次发现,执行长马尔可夫链的好处没有出现在该例子中。

表 27.4 缺失数据估算:对出现 25%的缺失数据与低相关性的线性回归,运用 MCMC 算法

	无数据缺失	成列删除	均值估算	马尔可夫链的长度			
				10	1 000	5 000	10 000
$\hat{\beta}_0$	1.121	1.162	1.142	1.149	1.155	1.154	1.141
	(0.099)	(0.130)	(0.103)	(0.104)	(0.103)	(0.104)	(0.101)
$\hat{\beta}_1$	1.099	0.930	1.052	1.026	1.020	1.004	1.044
	(0.107)	(0.134)	(0.121)	(0.127)	(0.128)	(0.124)	(0.124)
$\hat{\beta}_2$	1.102	1.122	1.215	1.130	1.157	1.137	1.151
	(0.107)	(0.134)	(0.124)	(0.128)	(0.129)	(0.129)	(0.119)
R^2	0.243	0.235	0.186				

27.8.2 *回归元出现缺失数据的 logit 回归*

我们再次考察,利用模拟数据的回归元出现缺失数据的非线性模型例子。在该模拟例子中,保持以前给定的数据生成过程,只是将因变量变成一种离散的二值变量。首先,对于线性回归例子,重新解释给定的模拟设计,因而 $y=y^*$ 潜变量。设数据生成过程是:

$$y_i^* = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i, \quad i=1,2,\cdots,N \tag{27.22}$$

于是,二值变量 y_i 依据下述规则生成:

$$y_i = \begin{cases} 1, & 当\ y_i^* > 0 \\ 0, & 当\ y_i^* \leqslant 0 \end{cases} \tag{27.23}$$

尽管数据生成过程是关于 probit 模型的,但我们将对 $y_i=0$ 的概率使用 logit 模型进行建模。如同 14.4.1 节所讨论的,logit 模型识别参数向量 $\boldsymbol{\beta}/\sigma$,其中,方差 $\sigma^2=$

$\pi^2/3$。就 β 的所有元素都被设置为 1 而言，logit 模型将给出近似−0.551的真实参数值的估计值。如同前面一样，用不提供信息的先验建立MCMC估计。

表 27.5 涵盖被人赞同的情况：出现 10%缺失数据，并且 x_1 与 x_2 之间有高度相关性，而表 27.6 涵盖不被人赞同的情况：出现 25%缺失数据，并且 x_1 与 x_2 之间有低相关性。

表 27.5 缺失数据估算：对出现 10%的缺失数据与高相关性的逻辑斯蒂回归估计，运用 MCMC 算法

				马尔可夫链的长度			
	无数据缺失	成列删除	均值估算	10	1 000	5 000	10 000
$\hat{\beta}_0$	−0.447	−0.498	−0.439	−0.527	−0.534	−0.531	−0.539
	(0.070)	(0.078)	(0.070)	(0.073)	(0.073)	(0.072)	(0.073)
$\hat{\beta}_1$	−0.597	−0.658	−0.602	−0.620	−0.673	−0.681	−0.675
	(0.096)	(0.108)	(0.098)	(0.106)	(0.102)	(0.101)	(0.103)
$\hat{\beta}_2$	−0.444	−0.474	−0.523	−0.597	−0.540	−0.536	−0.553
	(0.092)	(0.103)	(0.094)	(0.107)	(0.103)	(0.099)	(0.101)

表 27.6 缺失数据估算：对出现 25%的缺失数据与低相关性的逻辑斯蒂回归估计，运用 MCMC 算法

				马尔可夫链的长度			
	无数据缺失	成列删除	均值估算	10	1 000	5 000	10 000
$\hat{\beta}_0$	−0.447	−0.658	−0.582	−0.605	−0.609	−0.609	−0.599
	(0.070)	(0.097)	(0.070)	(0.074)	(0.074)	(0.073)	(0.076)
$\hat{\beta}_1$	−0.597	−0.434	−0.470	−0.447	−0.470	−0.471	−0.481
	(0.096)	(0.100)	(0.085)	(0.090)	(0.094)	(0.094)	(0.082)
$\hat{\beta}_2$	−0.444	−0.593	−0.648	−0.634	−0.615	−0.576	−0.596
	(0.092)	(0.108)	(0.089)	(0.084)	(0.086)	(0.086)	(0.094)

在第一种情况下，甚至就没有缺失数据而言，$\hat{\beta}_2$ 估计值偏离其期望值相当远。当马尔可夫链的长度从 10 增大到 1 000 时，点估计值稍微有变化。不过，进一步地当实施模拟时，点估计仅有一些微小的变化，我们将此结果解释成马尔可夫链收敛到其平稳分布的指示。

对于第二种例子，它涉及不被赞同的情况，其结果已由表 27.6 所示。主要差别在于，期望点估计与估计值之间的差异有点大于以前的情况。可是，更宽泛地讲，逻辑斯蒂回归的多重估算方法的效果类似于线性回归的效果。

27.9 应用研究

本章针对实际应用部分的分析表明，多重估算可能在理论上比估算数据更具优势。另外，基于模型的方法比那种机械方法比如均值估算或简单替补更缺少特设性。不过，在许多现实应用中，与上一节讨论的例子简单性相比，设计一种 MCMC 类型估算方法则可能是一项重要挑战。

在终端生成数据的多重估算与终端产出用于推断目标的估计系数之间，可能要画出一条界线。尽管这两种方法都将模型建立在第二种情形基础上，但这样做会涉及更为复杂的经济计量模型。布朗斯通和瓦莱塔(Brownstone and Valetta, 1996)，斯坦博林纳(Stinebrinkner, 1999)，肯尼克尔(Kennickell, 1998)，以及戴维、沙纳汉和谢弗(Davey, Shanahan, and Schafer, 2001)都给出了例子。

甚至最初目标是估算，缺乏广泛建模的问题可能不那么简单。例如，肯尼克尔在1995年的消费者财务调查研究中[肯尼克尔(Kennickell, 1998，第5页)]指出：

> 当调查包括相当多的变量时，就会出现大量的缺失或部分缺失(范围)信息，缺失信息的模式是高度异质性的，一些变量的分布出现高度偏斜，并且数据具有复杂结构，那么在出现缺失的条件下，对调查进行分析将是一项非常艰难的任务。此外，利用公开数据集合形式的任何人都会缺乏数据基准体系，而这被证明是认识缺失数据分布的一项重要因素。因此，即使基于纯效率的考量，对缺失数据进行设定的确是一件好事。

尽管问题具有复杂性，但肯尼克尔使用了类似于本章所讨论的那些估算方法。

斯坦博林纳(Stinebrinkner, 1999)同样面对下面缺失数据情形：成列删除“使得经济计量学家用极少的数据去估计关注模型”，为此，他探索一种两阶段基于似然模拟的方法，估计缺失数据的联合分布，并对首个教学时期的持续期限模型加以估计。

对于相对简单情况，可以运用像SAS软件的Proc程序包。而S-Plus与SOLAS也提供了软件支持。霍顿和利普希茨(Horton and Lipsitz, 2001)对计算机软件程序包给出一个有益的指南及综述。对于更多其他的信息，参见有关的web网站。

本章大多数分析均建立在假定具有一个可忽略缺失数据机制的基础上。从经济计量观点来看，这是一个重要的简化。例如，参见利拉德、史密斯和韦尔奇(Lillard, Smith, and Welch, 1986)，他们曾经评论估算缺失的人口普查简单替补方法。倘若缺失数据机制是不可忽略的，人们应该怎样继续做呢？在27.4节的符号下，不可忽略缺失数据机制蕴含，$\boldsymbol{\theta}$与ψ不是独立的。那么，人们必须用明确方式设定缺失数据机制，如同选择模型与损耗偏倚模型情况那样(参见第16章与23.5.2节)。谢弗(Schafer, 1997，第28页)给出了有关的参考文献。

27.10 文献注释

早期重要的参考文献包括利特尔和鲁宾(Little and Rubin, 1987)、鲁宾(Rubin, 1987)。阿利森(Allison, 2002)给出了一个相对非技术性又通俗易懂的缺失数据问题的介绍，以及参考文献。鲁宾(Rubin, 1996)则从历史观点出发，提供了一个综述。谢弗(Schafer, 1997)给出一个更完整的分析，涵盖分类数据、混合离散连续数据以及来自复杂调查的数据。

27.2 孟(Meng, 2000)针对缺失数据机制提出一种观点。

27.5 利特尔(Little, 1988, 1992)对线性回归含有缺失回归元的文献给出一个很好综述,其中既涵盖非基于模型方法,又涵盖基于模型方法。

习　题

27-1 考察任何线性或非线性回归模型,其中,因变量为 y,内生变量为 $\mathbf{x}$,还有 iid 误差 ε。证明,如果 $\mathbf{x}$ 出现缺失数据的概率与 y 无关,那么基于成列删除,该回归将给出条件均值函数的一致估计。[提示:证明给定 $\mathbf{x}$ 时 y 的条件分布没有受到缺失观测值的影响。]

27-2 [改编自古里耶克斯和蒙福特(Gouriéroux and Monfort, 1981)。]考察回归模型 $\mathbf{y}=\beta_1\mathbf{x}+\mathbf{Z}\boldsymbol{\beta}_2+\mathbf{u}$,其中,$\mathbf{y}$ 表示一个 $N\times1$ 维向量,$\mathbf{Z}$ 是一个 $N\times K$ 阶矩阵,$\mathbf{x}$ 是一个 $N\times1$ 维纯回归元向量,$\mathbf{x}$ 的某些元素出现缺失。假定观测值以随机方式出现缺失,且 $\mathrm{E}[\mathbf{u}|\mathbf{x},\mathbf{Z}]=\mathbf{0}$,同时 $\mathrm{E}[\mathbf{u}\mathbf{u}'|\mathbf{x},\mathbf{Z}]=\sigma^2\mathbf{I}_N$。不论是 $\mathbf{y}$ 还是 $\mathbf{Z}$ 均是完全观测的。提出下述方法用于处理缺失数据。假定 $\mathbf{x}$ 与 $\mathbf{Z}$ 有关的线性回归模型是 $\mathbf{x}=\mathbf{Z}\boldsymbol{\gamma}+\boldsymbol{\varepsilon}$,其中,$\mathrm{E}[\boldsymbol{\varepsilon}|\mathbf{Z}]=\mathbf{0}$,而 $\mathrm{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{Z}]=\sigma_\varepsilon^2\mathbf{I}_N$。于是,设 $\hat{\boldsymbol{\gamma}}=[\mathbf{Z}_c'\mathbf{Z}_c]^{-1}\mathbf{Z}_c'\mathbf{x}_c$,这里的下标“c”意指“完整数据”。估算值 $\hat{\mathbf{x}}_m=\mathbf{Z}_m[\mathbf{Z}_c'\mathbf{Z}_c]^{-1}\mathbf{Z}_c'\mathbf{x}_c$,其中,$\mathbf{x}_m$ 意指缺失观测值,$\mathbf{Z}_m$ 则是 $\mathbf{Z}$ 的相应值。从而,在用估算值代替 $\mathbf{x}$ 的缺失值之后,利用完整 N 个观测值集合,重新估计最初回归。

(a) 解释为什么基于完整观测值与估算观测值的 OLS 回归估计量都可能在有限样本时是有偏的?

(b) 需要什么样的额外条件可以证明,基于完整观测值加上估算值的 OLS 估计量是一致的?

(c) OLS 估计量是有效的吗?

27-3 考察下述观点:在数据估算之后,对模型进行估计,倘若对估算步骤不做调整,则估计准确性可能被夸大。换句话说,估算数据被看作生成变量,从而受限于 6.6 节曾讨论的序贯两步估计量问题。解释与缺失数据估算有关的调整是否在渐近形式上是必需的。

A 渐近理论

A.1 引 言

在附录中，我们考察当 $N\to\infty$ 时，随机变量序列（**sequence of random variables**）b_N 的特性。

在一些应用中，指标 N 表示样本量，而序列 b_N 表示估计量，比如 $\hat{\beta}$ 或 $\hat{\theta}$，或者是估计量的一个成分，就含有一个回归元的且没有截距的 OLS 而言，比如 $N^{-1}\sum_i x_i^2$ 或 $N^{-1}\sum_i x_i u_i$，或者是一个检验统计量。

对于估计理论来说，关注当 $N\to\infty$ 时序列的两方面特性就足够了。第一，我们考察 b_N 的依概率收敛（**convergence in probability**）到一个常值或者随机变量极限值 b，在下述将要定义的概率意义下这个常值或随机变量非常接近于 b_N。第二，如果极限值 b 是一个随机变量，该随机变量可能需要对原来序列进行重新标度，那么就要考察极限分布（**limit distribution**）。

通常，估计量是平均值（**averages**）或和（**sums**）的一个函数。于是，一种最容易的方法是，通过涉及平均特性的结果，即著名大数定律与中心极限定律来推导极限的一些结果。所用记号是平均值 $\bar{X}_N=N^{-1}\sum_i X_i$，其中，$X_i$ 表示对于随机变量作为平均的一般记号，而对于用 $\mathbf{x}_i$ 表示回归元向量的情况来说，X_i 不应与之混淆。例如，就含有单个回归元且没有截距的 OLS 而言，我们将大数定律用到 $X_i=x_i^2$ 的平均上，而将中心极限定律用到 $X_i=x_i u_i$ 平均上。

表 A.1 概括出附录余下部分所要表述的定义与定理。这些内容都没有给出证明，却给出某种讨论。关注内容是，通常人们使用横截面数据时为获得渐近正态估计量而使用的一些结果。另一些结果满足运用非参数估计的需要，当数据依赖于参数时，满足运用参数进行估计所需的结果，以及当数据具有单位根时，运用时间序列进行估计所需的结果。

第一个重要的概念是 A.2 节表述的依概念收敛。这是利用 A.3 节给出的大数定律建立起来的。另一个重要概念是 A.4 节表述的依分布收敛。收敛到正态分布可以利用 A.5 节给出的中心极限定律建立。对于多元正态分布来说，更进一步的结果及常用术语已在 A.6 节给出。A.7 节表述了渐近分析中通常广泛使用的简便记号，即随机数量级。A.8 节阐述期望的某些有用性质。

表 A.1 渐近理论:定义与定理

定义	定理	名称	式子
A. 1		依概率收敛	(A. 1)
A. 2		一致性	(A. 2)
	A. 3	斯卢茨基	(A. 3)
A. 4		均方收敛	(A. 4)
	A. 5	切比雪夫不等式	(A. 5)
A. 6		几乎处处收敛	(A. 6)
A. 7		大数定理	(A. 7)
	A. 8	柯尔莫哥洛夫 LLN	
	A. 9	马尔可夫 LLN	
A. 10		依分布收敛	(A. 9)
	A. 11	连续映射	(A. 10)
	A. 12	变 换	(A. 11)
A. 13		中心极限定理	(A. 13)
	A. 14	林德伯格—莱维 CLN	
	A. 15	李雅普诺夫 CLT	
	A. 16	克莱姆—沃尔德方法	
	A. 17	正态极限乘积法则	(A. 15)
A. 18		渐近分布	(A. 17)
A. 19		渐近方差	(A. 18)
A. 20		估计渐近方差	(A. 19)
A. 21		渐近有效性	
A. 22		随机数量级	

A.2 依概率收敛

由于样本固有的随机性,尽管该样本可以无限大,但我们永远不能确定诸如估计量 $\hat{\theta}$(经常表示成 $\hat{\theta}_N$,以表明它是一个序列)的序列 b_N 位于给定其极限的某个很小距离之内。不过,我们能大致如此确定。利用各种不同方式表述这种接近于确定性的形式,以此对应于随机变量序列收敛到其极限的不同类型。经济计量学最广泛运用的收敛极限类型是依概率收敛。

A.2.1 *依概率收敛*

回顾,非随机实数序列 $\{a_N\}$ 收敛到 a,如果对于任意的 $\varepsilon>0$,存在 $N^*=N^*(\varepsilon)$,使得对于所有 $N>N^*$,有:

$$|a_N-a|<\varepsilon$$

例如,当 $a_N=2+3/N$ 时,其极限是 $a_N=2$,这是因为 $|a_N-a|=|2+3/N-2|=|3/N|<\varepsilon$,对于所有 $N>N^*=3/\varepsilon$。

更一般地讲,当我们拥有随机变量序列时,因其固有的随机性,甚至对于很大

的 N，我们都不能确定其极限的某一个范围。相反，我们需要位于 ε 某范围内的概率可以是任意接近于 1 的。因而，我们要求：

$$\lim_{N\to\infty}\Pr[|b_N-b|<\varepsilon]=1$$

对于任意的 $\varepsilon>0$ 。正式定义如下：

定义 A.1(依概率收敛)：一个随机变量序列 $\{b_N\}$ 依概率收敛到 b，如果对于任意的 $\varepsilon>0$ 且 $\delta>0$，存在 $N^*=N^*(\varepsilon,\delta)$，使得对于所有 $N>N^*$，有：

$$\Pr[|b_N-b|<\varepsilon]=1-\delta \tag{A.1}$$

我们将其写成 $\text{plim}\, b_N=b$，其中，plim 表示概率极限简略写法，或 $b_N\xrightarrow{p}b$。

注意到，b 可能是一个常值或随机变量。对向量随机变量的推广，比如参数向量估计量，可以直接进行。依概率收敛包括了作为特殊情况的实变量序列收敛的通常定义。定义 A.1 是对纯量随机变量序列来定义的。我们对 $\mathbf{b}_N$ 的每一个元素应用理论，或者用纯量 $(\mathbf{b}_N-\mathbf{b})'(\mathbf{b}_N-\mathbf{b})=(b_{1N}-b_1)^2+\cdots+(b_{KN}-b_K)^2$ 或其平方根 $\|\mathbf{b}_N-\mathbf{b}\|$ 来代替 $|b_N-b|$。

当序列 $\{\mathbf{b}_N\}$ 作为参数估计值 $\hat{\boldsymbol{\theta}}$ 的序列时，我们有下述大样本无偏性的类似形式。

定义 A.2(一致性)：估计量 $\hat{\boldsymbol{\theta}}$ 关于 $\boldsymbol{\theta}_0$ 是一致的，如果：

$$\text{plim}\,\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}_0 \tag{A.2}$$

$\boldsymbol{\theta}_0$ 的下标“0”已在 5.2.3 节给出解释。注意到，无偏性并不要求蕴含着一致性。无偏性仅仅表明，$\hat{\boldsymbol{\theta}}$ 的期望值是 $\boldsymbol{\theta}_0$，而且它允许在 $\boldsymbol{\theta}_0$ 附近出现变异性，这一点没有随样本量趋向于无穷而消失。同样地，一致估计量并不要求无偏性。例如，将 $1/N$ 添加到一个无偏且一致估计量上，会产生一个新的有偏估计量，但仍是一致的。

尽管向量随机变量 $\{\mathbf{b}_N\}$ 的序列可以收敛到一个随机变量 $\mathbf{b}$，但在许多经济计量应用中，$\{\mathbf{b}_N\}$ 收敛到一个常值。例如，我们希望参数的一个估计量依概率收敛到参数自身。应该注意，一些结果只有当极限值 $\mathbf{b}$ 是常值时才可以应用。

定理 A.3(斯卢茨基定理)：设 $\mathbf{b}_N$ 是一个有限维的随机变量向量，而 $g(\cdot)$ 表示在常值向量点 $\mathbf{b}$ 处是连续的一个实值函数，于是：

$$\mathbf{b}_N\xrightarrow{p}\mathbf{b}\Rightarrow g(\mathbf{b}_N)\xrightarrow{p}g(\mathbf{b}) \tag{A.3}$$

雨宫(Amemiya，1985，第 89 页[1])给出了此定理的证明。鲁德(Ruud，2000)阐明了有关结果，还可参见拉奥(Rao，1973，第 124 页)，即设极限 $\mathbf{b}$ 是一个极限向量，却是以将 $g(\cdot)$ 约束为处处连续为代价的。注意到，一些作者反而将下面的定理 A.12 称为斯卢茨基定理。

定理 A.3 是导致经济计量学中相对于有限样本结果而言渐近结果盛行的重要原因之一。它表述了一种非常方便的性质，但该性质对于期望值却不成立。例

〔1〕 原著中这里为第 79 页，应为第 89 页。——译者注

如，$\text{plim}(b_{1N}, b_{2N})=(b_1, b_2)$ 蕴含着 $\text{plim}(b_{1N}b_{2N})=(b_1b_2)$，而通常 $E[b_{1N}b_{2N}]$ 与 $E[b_1]E[b_2]$ 却是不同的。

A.2.2 其他收敛方式

通常，比较容易建立一些其他的收敛方式，反过来也蕴含着依概率收敛。

为了完整起见，逐一给出这些其他方式。通常人们广泛使用下一节给出的大数定律。

定义 A.4(均方收敛)：一个随机变量序列 $\{b_N\}$ 称为依均方收敛到随机变量 b，如果：

$$\lim_{N\to\infty} E[(b_N-b)^2]=0 \tag{A.4}$$

我们将其写成 $b_N \xrightarrow{m} b$。因为 $b_N \xrightarrow{m} b$ 蕴含着 $b_N \xrightarrow{p} b$[参见拉奥(Rao，1973，110 页)]，故依均方收敛十分有用，而且很容易证明这一点。然而，这确实要求 b_N 方差存在。如果 $E[b_N]=b$，那么需要证明，当 $N\to\infty$ 时，b_N 方差趋于 0。反之，如果 b_N 对 b 来说是有偏的，那么要求方差之和及偏倚平方趋于 0。

经常用于证明依概率收敛的另一个结果是切比雪夫不等式。

定理 A.5(切比雪夫不等式)：对于任何一个随机变量 Z，其均值为 μ 且方差为 σ^2，则：

$$\Pr[(Z-\mu)^2>k]\leqslant\sigma^2/k, \quad \text{对于任何 } k>0 \tag{A.5}$$

对它的证明，可参见[拉奥(Rao，1973，第 95 页)]。广义切比雪夫不等式是用任意非负函数 $g(Z)$ 代替定理 A.5 中的 $(Z-\mu)^2$，然后证明 $\Pr[g(Z)>k]\leqslant E[g(Z)]/k$，对于任何 $k>0$。参见雨宫(Amemiya，1985，第 87 页)。

定理 A.5 经常通过用 b_N 代替 Z 来证实依概率收敛。该定理要求 b_N 的均值与方差均存在，这很容易从涉及独立随机变量平均的估计量中获得。不过，在这种情况下，我们经常采用甚至更容易的路线，并直接将大数定律用到平均值上得到概率极限。

从概念上讲，更困难的收敛类型是几乎必然收敛。

定义 A.6(几乎必然收敛)：一个随机变量序列 $\{b_N\}$ 称为几乎必然收敛(**converge almost surely**)到 b，如果：

$$\Pr[\lim_{N\to\infty} b_N=b]=1^{[1]} \tag{A.6}$$

这被记为 $b_N \xrightarrow{as} b$。几乎必然收敛蕴含着依概率收敛[参见拉奥(Rao，1973，第 111 页)]。依概率收敛比几乎必然收敛允许 b_N 中出现的特性更为不规则。

对于 b 来说，几乎必然收敛还称为强一致性(**strong consistence**)，而与之相区别，依概率收敛称 b 为弱一致性(**weak consistence**)。依概率收敛比较容易理解，而且对于大多数经济计量应用来说，这已足够了。

---

[1] 原著中该式右端缺少"＝1"，现已加上。——译者注

A.3 大数定律

大数定律是依概率收敛(或几乎必然收敛)的特殊情况,此时,序列$\{b_N\}$是样本平均值,即$b_N=\overline{X}_N$,其中:

$$\overline{X}_N=\frac{1}{N}\sum_{i=1}^{N}X_i \tag{A.7}$$

注意到,这里,X_i 表示随机变量的一般记号,而在回归背景下,它不一定表示回归元变量。

大数定律与运用(A.1)给出的定义(δ,ε)的笨拙选择方法或蕴含着依概率收敛的其他方式相比,都更容易建立序列$\{b_N\}$的概率极限。

定义 A.7(大数定律): 弱大数定律(**weak law of large numbers**, 记为 LLN)是在

$$(\overline{X}_N-\mathrm{E}[\overline{X}_N])\xrightarrow{p}0 \tag{A.8}$$

的条件下,规定了 $\overline{X}_N$ 中各个 X_i 项的状况。

然而,对于强大数定律来说,收敛就是几乎必然收敛。

考虑将大数定律建立为 $\overline{X}_N$ 趋于其期望值是有益的,尽管严格地讲,它蕴含着 $\overline{X}_N$ 趋于其期望值极限的比较弱条件,因为式(A.8)蕴含着:

$$\operatorname{plim}\overline{X}_N=\lim \mathrm{E}[\overline{X}_N]$$

若 X_i 具有共同均值μ ,则这简化成 $\operatorname{plim}\overline{X}_N=\mu$。

大数定律的两个重要例子如下:

定理 A.8(柯尔莫哥洛夫 LLN): 设$\{X_i\}$是 iid(独立同分布)的,当且仅当 $\mathrm{E}[X_i]=\mu$ 存在且 $\mathrm{E}[|X_i|]<\infty$,则$(\overline{X}_N-\mathrm{E}[\overline{X}_N])\xrightarrow{as}0$。

定理 A.9(马尔可夫 LLN): 设$\{X_i\}$是 inid(独立但非同分布)的。满足$\mathrm{E}[X_i]=\mu_i$ 且 $\mathrm{V}[X_i]=\sigma_i^2$。如果$\sum_{i=1}^{\infty}(\mathrm{E}[|X_i-\mu_i|^{1+\delta}]/i^{1+\delta})<\infty$,对于某个 $\delta>0$,则$(\overline{X}_N-\mathrm{E}[\overline{X}_N])\xrightarrow{as}0$。

参见怀特(White, 2001a,第 32 页与第 35 页)对这些定理的叙述,以及拉奥(Rao, 1973,第 114~116 页)的证明。两个定律为我们提供了几乎必然收敛的比较强的结果,这蕴含着人们想要得到的依概率收敛。拉奥(Rao, 1973)把定理 A.8 称为柯尔莫哥洛夫第 2 大数定理(LLN2),并对于特殊情况 $\delta=1$ 表述了定理 A.9,他称之为柯尔莫哥洛夫第 1 大数定理(LLN1)。

柯尔莫哥洛夫大数定律允许 X_i 的方差甚至不存在,却以要求同分布作为代价。它简化成 $\overline{X}_N\xrightarrow{as}\mu$,其中,$\mu=\mathrm{E}[X]$。该定律的一个弱形式是辛钦[1](Khinchine)定理,它可表述成:对于 iid $\{X_i\}$来说,E[X]的存在蕴含着依概率收

[1] 辛钦(Khinchine, 1894~1959 年),苏联数学家。——译者注

敛，这对大多数经济计量应用来说足够了。

马尔可夫大数定律不再要求同分布，但它需要大于一阶的绝对矩存在。δ 的一个明显选择是 $\delta=1$。于是，需要方差存在，而且边条件是 $\sum_{i=1}^{\infty}(\sigma_i^2/i^2)<\infty$。该方差可以变化，甚至随 i 而增大，假若它增大得并不快，(σ_i^2/i^2) 将具有无穷和。当 $\sigma_i^2=\sigma^2$ 时，由于 $\sum_{i=1}^{\infty}1/i^2$ 收敛，所以边条件得以满足，当 $\sigma_i^2=i\sigma^2$ 时，因为 $\sum_{i=1}^{\infty}1/i$ 发散，故边条件不满足。

在大部分微观经济计量学的应用中，包括含有分层抽样或固定回归元的回归，需要更为复杂的马尔可夫大数定律。

大数定律颇为引人注目，原因在于这些定理要求个体成分 X_i 的假设，而不是平均值序列 $\bar{X}_N$ 的假设。由于大部分估计量及检验统计量都是数据的平均值与不可观测随机变量的函数，所以大数定律是经济计量学家用于证明依概率收敛的一种最普遍方法。

A.4 依分布收敛

给定一致性，估计量 $\hat{\theta}$ 具有退化分布[1]，即当 $N\rightarrow\infty$ 时在 θ_0 处出现常值。我们需要放大或者重新标度 $\hat{\theta}$，以便获得当 $N\rightarrow\infty$ 时具有非退化分布的随机变量。一种适当的标度因子通常是 $\sqrt{N}$，在此情况下，我们考察随机变量序列 $b_N=\sqrt{N}(\hat{\theta}-\theta_0)$ 的特性。

通常，序列 b_N 的第 N 个随机变量具有极端复杂的累积分布函数(cdf) F_N。如同任何其他函数 F_N 一样，这可能具有极限函数，其中的收敛是在通常数学意义下的。

定义 A.10(依分布收敛)：一个随机变量序列 $\{b_N\}$ 称为依分布收敛到随机变量序列 b，如果在 F 的每一个连续性点上：

$$\lim_{N\rightarrow\infty}F_N=F \tag{A.9}$$

其中，F_N 表示 b_N 的分布，F 表示 b 的分布，而收敛是在通常数学意义下的。

我们将其写成 $b_N\xrightarrow{d}b$，并称 F 为 $\{b_N\}$ 的极限分布。

---

[1] 退化分布(degenerate distribution)：n 个随机变量 $X_1,\cdots,X_n$ 的联合分布被称为退化的，如果这 n 个变量之间至少有一种关系 $g(X_1,\cdots,X_n)=0$ 以概率 1 成立，对于所有 $X_1,\cdots,X_n$ 来说，函数 $g(\cdot)$ 本身不是恒等常值函数。

在单个随机变量 X 的情况下，有：

$$P(X=a)=1$$

其对应的 cdf 是：

$$P(X\leqslant x)=F(x)=\begin{cases}0, & x<a\\ 1, & x\geqslant a\end{cases}$$

而特征函数是 $\phi(t)=e^{ita}$。此分布的矩是 $\mu_k'=E(X^k)=a^k$，$k=1,2,\cdots$，而 $\text{var}(X)=0$。有时候，人们概括地说，这个退化分布便是描述"非随机变量"。其逆命题同样成立。倘若某个随机变量 X 具有有限期望且零方差，则 $P(X=E[X])=1$。

依概率收敛蕴含着依分布收敛：即 $b_N \xrightarrow{p} b$ 蕴含着 $b_N \xrightarrow{d} b$[参见拉奥(Rao，1973，第 122 页)]。

一般地讲，其逆不成立。例如，设 $b_N = X_N$，X 的第 N 个实现值～$\mathcal{N}[\mu, \sigma^2]$。于是，$b_N \xrightarrow{d} b \sim \mathcal{N}[\mu, \sigma^2]$，可是很明显，$(b - b_N)$ 具有当 $N \to \infty$ 时并不消失的方差，因此，b_N 不会依概率收敛到 b。

然而，在 b 为常值的特殊情况下，$b_N \xrightarrow{d} b$ 蕴含着 $b_N \xrightarrow{p} b$[参见拉奥(Rao，1973，第 120 页)]。在此情况下，该极限分布是退化的，其所有质量都位于 b 处。

为将极限分布推广到向量随机变量，可直接定义 F_N 与 F 分别是向量 $\mathbf{b}_N$ 与 $\mathbf{b}$ 的各自 cdf。

定理 A. 11(连续映射定理)：设 $\mathbf{b}_N$ 是一个有限维的随机变量向量，并设 $g(\cdot)$ 表示连续实值函数。于是：

$$\mathbf{b}_N \xrightarrow{d} b \Rightarrow g(\mathbf{b}_N) \xrightarrow{d} g(\mathbf{b}) \tag{A. 10}$$

其具体证明，参见拉奥(Rao，1973，第 124 页)。定理 A. 11 是依分布收敛的，这类似于依概率收敛的定理 A. 3。

下述定理考虑了通过将一个具有极限分布的序列加上或乘以或除以一个依概率收敛到常值的序列所具有的变换效果。

定理 A. 12(变换定理)：如果 $a_N \xrightarrow{d} a$ 且 $b_N \xrightarrow{p} b$，其中，a 表示一个随机变量，b 表示一个常值，那么：

(i) $a_N + b_N \xrightarrow{d} a + b$

(ii) $a_N b_N \xrightarrow{d} ab$ (A. 11)

(iii) 倘若 $\Pr[b=0]=0$，$a_N / b_N \xrightarrow{d} a/b$

其证明参见拉奥(Rao，1973，第 122 页)。定理 A. 12 还称为克拉默定理。它也称为斯卢茨基定理，该名称我们已应用于定理 A. 3。

定理 A. 12 特别有用，因为它允许人们分别求出 a_N 的极限分布与 b_N 的概率极限，而不用考察 a_N 与 b_N 的联合特性。结论(ii)尤其有用，而且它有时被称为乘积法则。

A.5 中心极限定理

当序列$\{b_N\}$是样本平均值时，中心极限定理就是依分布收敛的定理。中心极限定理提供了比使用可供选择的其他方法诸如笨拙的式(A. 9)更为简单地获得序列$\{b_N\}$极限分布的方法。

由大数定律知，样本均值具有退化分布，因为它收敛到一个常值即 $\lim \mathrm{E}[\bar{X}_N]$ 上。因此，我们借助于它的标准差标度$(\bar{X}_N - \mathrm{E}[\bar{X}_N])$，构造一个具有单位方差的随机变量，该随机变量可以收敛到一个非退化分布。

定理 A.13(中心极限定理):设:

$$Z_N=\frac{\bar{X}_N-\mathrm{E}[\bar{X}_N]}{\sqrt{\mathrm{V}[\bar{X}_N]}} \tag{A.12}$$

其中,$\bar{X}_N$ 表示样本均值。中心极限定律(**central limit theorem**,记为 CLT)在

$$Z_N \xrightarrow{d} \mathcal{N}[0,1] \tag{A.13}$$

条件下,即 Z_N 依分布收敛到标准正态随机变量的条件下,规定了 $\bar{X}_N$ 中各个 X_i 项的状况。

通过构造知道,Z_N 具有均值为 0 且方差为 1,因此,需要证明的内容是正态性。中心极限定理的正式证明,可以通过获得 Z_N 的特征函数,即广义矩母函数[1],并且证明,当 $N\to\infty$ 时,它收敛到标准正态分布的特征函数。

注意到,如果 $\bar{X}_N$ 满足中心极限定理,那么关于函数 $h(\cdot)$ 比如 $h(N)=\sqrt{N}$,$h(N)\bar{X}_N$ 也满足中心极限定理,因为:

$$Z_N=\frac{h(N)\bar{X}_N-\mathrm{E}[h(N)\bar{X}_N]}{\sqrt{\mathrm{V}[h(N)\bar{X}_N]}}$$

在许多应用中,将中心极限定理用于正规化 $\sqrt{N}\bar{X}_N=N^{-1/2}\sum_{i=1}^{N}X_i$ 上是方便的,因为 $\mathrm{V}[\sqrt{N}\bar{X}_N]$ 是有限的。

中心极限定理的例子包括:

定理 A.14(林德贝格—勒维 CLT):设 $\{X_i\}$ 是 iid 的,满足 $\mathrm{E}[X_i]=\mu$ 且 $\mathrm{V}[X_i]=\sigma^2$。那么,$Z_N \xrightarrow{d} \mathcal{N}[0,1]$。

对该定理的证明,参见拉奥(Rao, 1973,第 127 页)。

这是通常统计学导论出现的中心极限定理,在 *iid* 情况下,它十分有用。由于 X_i 是 iid $[0,\sigma^2]$,所以 Z_N 可简化成更熟悉的:

$$Z_N=\frac{\bar{X}_N-\mu}{\sigma/\sqrt{N}}$$

注意到,在 iid 情况下,唯一要求 μ 存在,以此可确保 $\bar{X}_N \xrightarrow{p} \mu$,而要获得极限正态分布,则需要额外假设,即 σ^2 存在。

在诸如含有固定回归元的 OLS 一些应用中,iid 假设是不恰当的。人们能应用关于 $\{X_i\}$ inid 的中心极限定理,尽管这需要做出额外的假设。

定理 A.15(李雅普诺夫 CLT):设 $\{X_i\}$ 是独立的,满足 $\mathrm{E}[X_i]=\mu_i$ 且 $\mathrm{V}[X_i]=\sigma_i^2$。如果 $\lim(\sum_{i=1}^{N}\mathrm{E}[|X_i-\mu_i|^{2+\delta}])/(\sum_{i=1}^{N}\sigma_i^2)^{(2+\delta)/2}=0$,对于某个选定的 $\delta>0$,那么 $Z_N \xrightarrow{d} \mathcal{N}[0,1]$。

李雅普诺夫中心极限定理的这种变形,已经由怀特(White, 2001a,第 119 页)证明。拉奥(Rao, 1973,第 128 页)阐述了 $\delta=1$ 的特殊情况。

---

[1] 又称为广义矩生成函数。——译者注

李雅普诺夫中心极限定理的主要附加假设是，高于二阶绝对矩存在。还要注意，附加假设与 iid 数据对应的 LLN 的比较。对于 inid X_i 而言：

$$Z_N=\frac{\sum_{i=1}^{N}X_i-\sum_{i=1}^{N}\mu_i}{\sqrt{\sum_{i=1}^{N}\sigma_i^2}}$$

定理 A. 14 与 A. 15 都是更一般的林德贝格—费勒中心极限定理的特殊情况[参见拉奥(Rao，1973，第 128 页)]。林德贝格—费勒中心极限定理具有很难验证的边条件。

在大部分微观经济计量学中，包括含有分层抽样或固定回归元的回归，经常运用更为复杂的李雅普诺夫中心极限定理。

A.6 多元正态极限分布

本节关注一般的带有多元正态极限分布估计量的微观经济计量应用情况。

A. 6. 1 *多元正态极限分布*

前面已阐述的中心极限定理是有关纯随机变量序列的情况。利用下述结果，可将中心极限定理推广到向量随机变量序列上。

定理 A. 16(克拉默—沃尔德方法)：设$\{\mathbf{b}_N\}$是 $k\times1$ 维随机向量的序列。如果 $\boldsymbol{\lambda}'\mathbf{b}_N$ 对于每一个 $k\times1$ 维常值非零向量 $\boldsymbol{\lambda}$，都收敛到一个正态随机变量上，那么 $\mathbf{b}_N$ 收敛到多元正态随机变量。

拉奥(Rao，1973，第 128 页)提供了并不局限于正态分布的更一般结果。

该结果的优点是，若 $\mathbf{b}_N$ 是一个平均向量，则 $\boldsymbol{\lambda}'\mathbf{b}_N=\lambda_1 b_{1N}+\cdots+\lambda_k b_{kN}$ 是一个纯量平均值，而且我们能应用前面一节给出的纯量中心极限定理。从而，得到：

$$\frac{\boldsymbol{\lambda}'\mathbf{b}_N-\boldsymbol{\lambda}'\boldsymbol{\mu}_N}{\sqrt{\boldsymbol{\lambda}'\mathbf{V}_N\boldsymbol{\lambda}}}\xrightarrow{d}\mathcal{N}[0,1]$$

其中，$\boldsymbol{\mu}_N=\mathrm{E}[\mathbf{b}_N]$，$\mathbf{V}_N=\mathbf{V}[\mathbf{b}_N]$，在此情况下，得出如下结论：

$$\mathbf{V}_N^{-1/2}(\mathbf{b}_N-\boldsymbol{\mu}_N)\xrightarrow{d}\mathcal{N}[\mathbf{0},\mathbf{I}] \tag{A. 14}$$

这一结果将在 A. 6. 3 节进一步解释。

A. 6. 2 *线性变换*

微观经济计量运用的估计量经常表述成 $\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)=\mathbf{H}_N\mathbf{a}_N$，其中，plim $\mathbf{H}_N$ 存在且 $\mathbf{a}_N$ 服从极限正态分布。这个积的分布或者 a_N 的线性变换，可从定理 A. 12(变换定理)的(ii)部分直接得到。我们可用一种形式重新叙述它，该形式可得出许多估计量。

定理 A. 17(极限正态积准则)：如果向量 $\mathbf{a}_N\xrightarrow{d}\mathcal{N}[\boldsymbol{\mu},\mathbf{A}]$，并且矩阵 $\mathbf{H}_N\xrightarrow{p}\mathbf{H}$，其中，$\mathbf{H}$ 表示正定的，那么：

$$\mathbf{H}_N\mathbf{a}_N \xrightarrow{d} \mathcal{N}[\mathbf{H}\boldsymbol{\mu}, \mathbf{HAH}'] \tag{A.15}$$

定理 A. 17 能直接应用到估计量上。例如,将 OLS 估计量

$$\sqrt{N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) = \left(\frac{1}{N}\mathbf{X}'\mathbf{X}\right)^{-1}\frac{1}{\sqrt{N}}\mathbf{X}'\mathbf{u}$$

处理成为 $\mathbf{H}_N=(N^{-1}\mathbf{X}'\mathbf{X})^{-1}$ 与 $\mathbf{a}_N=N^{-1/2}\mathbf{X}'\mathbf{u}$ 的乘积,从而我们求出 $\mathbf{H}_N$ 的 plim 以及 $\mathbf{a}_N$ 的极限分布。

定理 A. 17 还可用于证明,通过极限分布不变的一致估计量替换极限分布方差是正确的。如果已经证明:

$$\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}[\mathbf{0},\mathbf{B}]$$

那么由定理 A. 17,可得:

$$\mathbf{B}_N^{-1/2}\times\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}[\mathbf{0},\mathbf{I}]$$

对于任何一个 $\mathbf{B}$ 对而言,$\mathbf{B}_N$ 都是一致估计值且是正定的。

A. 6. 3 *极限方差矩阵*

从记号上看,正式的多元变量中心极限定理会产生繁琐的结果形式,比如式(A. 14)。一旦用 $\mathbf{V}_N^{-1/2}$ 左乘,并应用定理 A. 17,我们可用简单形式重新表述成:

$$\mathbf{b}_N-\boldsymbol{\mu}_N \xrightarrow{d} \mathcal{N}[\mathbf{0},\mathbf{V}]$$

其中,$\mathbf{V}=\text{plim}\,\mathbf{V}_N$,并假定 $\mathbf{b}_N$ 与 $\mathbf{V}_N$ 被适当地标度,以使 $\mathbf{V}$ 存在且是正定的。

许多作者都以各种不同方式来表述极限方差矩阵 **V(limt variance matrix)。**

一般定义是:

$$\mathbf{V}=\text{plim}\,\mathbf{V}_N$$

这是一种最普遍的表述结果方式,而且本书就是使用这种形式。在固定回归元的情况下,它可简化成 $\mathbf{V}=\lim\mathbf{V}_N$。

在微观经济计量学一些估计例子里,矩阵 $\mathbf{V}_N$ 经常是矩阵平均值,比如说:

$$\mathbf{V}_N = \frac{1}{N}\sum_{i=1}^{N}\mathbf{S}_i$$

其中,$\mathbf{S}_i$ 表示方阵,它是第 i 个观测值的参数与数据的函数。给定对于不同 i 的独立性,通常可应用大数定律,从而得出 $\mathbf{V}_N-\mathrm{E}[\mathbf{V}_N]\xrightarrow{d}\mathbf{0}$。于是:

$$\mathbf{V} = \lim \mathrm{E}[\mathbf{V}_N] = \lim\frac{1}{N}\sum_{i=1}^{N}\mathrm{E}[\mathbf{S}_i]$$

这是雨宫(Amemiya, 1985)曾经使用的表述形式。

若 $\mathbf{S}_i$ 是 iid 的 ,则对于所有观测值来说,$\mathrm{E}[\mathbf{S}_i]=\mathrm{E}[\mathbf{S}]$,因而,简单随机抽样导致了比较简单的表达式:

$$\mathbf{V}=\mathrm{E}[\mathbf{S}]$$

例如，纽韦和麦克法登(Newey and McFadden, 1994)与伍德里奇(Wooldridge, 2002)都曾用过该形式。

举一个例子，考察带有同方差误差的 OLS 估计量，因此，$\sqrt{N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\xrightarrow{d}\mathcal{N}[\mathbf{0},\sigma^2\mathbf{M}_{\mathbf{xx}}^{-1}]$。于是，如果应用研究大数定律，$\mathbf{M}_{\mathbf{xx}}=\mathrm{plim}\,N^{-1}\sum_i\mathbf{x}_i\mathbf{x}_i'$ 能被重新表述成 $\mathbf{M}_{\mathbf{xx}}=\lim N^{-1}\sum_i\mathrm{E}[\mathbf{x}_i\mathbf{x}_i']$，而在简单随机抽样下，则可表述成 $\mathbf{M}_{\mathbf{xx}}=\mathrm{E}[\mathbf{xx}']$。

人们也会得到 $\mathbf{V}$ 的更为复杂形式，比如三明治形式 $\mathbf{ABA}'$。于是，前面讨论可应用到每一个元素上。例如，在随机抽样下，若 $\mathbf{B}=N^{-1}\sum_i\mathbf{S}_i$，则 $\mathbf{B}=\mathrm{plim}\,\mathbf{B}_N$ 可以表述成 $\mathbf{B}=\lim\mathrm{E}[\mathbf{B}_N]$ 或者 $\mathbf{B}=\mathrm{E}[\mathbf{S}]$。

A.6.4 *渐近分布与方差*

为了获得估计量的极限分布，由于理论原因，我们以序列 $b_N=\sqrt{N}(\hat{\theta}-\boldsymbol{\theta}_0)$ 进行分析，以此确保当 $N\rightarrow\infty$ 时，b_N 有非零方差。于是，b_N 的极限分布是正态分布，而且许多作者都说 b_N 服从渐近正态的，并将极限方差矩阵称为 b_N 的渐近方差。

运用 $\hat{\boldsymbol{\theta}}$ 自身的分布及方差矩阵重新表述结果是十分方便的。

定义 A.18($\hat{\boldsymbol{\theta}}$ 的渐近分布)：如果：

$$\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\xrightarrow{d}\mathcal{N}[\mathbf{0},\mathbf{B}] \tag{A.16}$$

也就是说，$\hat{\boldsymbol{\theta}}$ 在大样本下服从渐近正态分布，满足：

$$\hat{\boldsymbol{\theta}}\sim\mathcal{N}[\boldsymbol{\theta}_0,N^{-1}\mathbf{B}] \tag{A.17}$$

其中，"在大样本下"这一术语意味着，N 对于式(A.16)成为良好近似来说是充分大的，但并没有大到使式(A.17)的方差趋于 0。

结果(A.17)可由式(A.16)得出，因为用 $\sqrt{N}$ 除以随机变量导致了用 N 除以其方差。

缩写记号隐含地假定了渐近正态性，并使用下面术语。

定义 A.19($\hat{\boldsymbol{\theta}}$ 的渐近方差)：如果式(A.16)成立，我们称 $\hat{\boldsymbol{\theta}}$ 的渐近方差矩阵是：

$$\mathrm{V}[\hat{\boldsymbol{\theta}}]=N^{-1}\mathbf{B} \tag{A.18}$$

定义 A.20($\hat{\boldsymbol{\theta}}$ 的估计渐近方差)：如果式(A.16)成立，我们称 $\hat{\boldsymbol{\theta}}$ 的估计渐近方差是：

$$\hat{\mathrm{V}}[\hat{\boldsymbol{\theta}}]=N^{-1}\hat{\mathbf{B}} \tag{A.19}$$

其中，$\hat{\mathbf{B}}$ 表示 $\mathbf{B}$ 的一致估计值。

一些作者在定义 A.19 与 A.20 中使用 $\mathrm{Avar}[\hat{\boldsymbol{\theta}}]$ 与 $\widehat{\mathrm{Avar}}[\hat{\boldsymbol{\theta}}]$，以避免潜在地与方差算子符号 $\mathrm{V}[\cdot]$ 相混淆。很明显，这里的 $\mathrm{V}[\hat{\boldsymbol{\theta}}]$ 意味着估计量的渐近方差，因为本书中一些估计量具有有限样本方差的闭形式表达式。

举一个定义 A.18～A.20 的例子，如果 $\{X_i\}$ 是 iid 的 $[\mu,\sigma^2]$，那么由林德贝格—勒维中心极限定理得出，$\sqrt{N}(\bar{X}_N-\mu)/\sigma\xrightarrow{d}\mathcal{N}[0,1]$，或者等价地 $\sqrt{N}\bar{X}_N\xrightarrow{d}$

$\mathcal{N}[0,\sigma^2]$[1]。我们就说，在渐近形式上 $\overline{X}_N \sim \mathcal{N}[\mu,\sigma^2/N]$；$\overline{X}_N$ 的渐近方差是 σ^2/N；$\overline{X}_N$ 的估计渐近方差是s^2/N，其中，s^2 表示 σ^2 的一致估计量，比如 $s^2=\sum_i(X_i-\overline{X}_N)^2/(N-1)$。

A.6.5 *渐近效率*

在有限样本中，无偏估计的方差协方差矩阵的克拉默—拉奥下界是 $-(\mathrm{E}[\partial^2 \ln \mathrm{L}_N/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'|_{\boldsymbol{\theta}_0}])^{-1}$。该结果可被推广到作为渐近正态的一致估计量上。

定义 A.21(渐近效率)：$\boldsymbol{\theta}$ 的一致渐近正态估计量 $\hat{\boldsymbol{\theta}}$ 称为是渐近有效的，如果 $\hat{\boldsymbol{\theta}}$ 具有等于克拉默—拉奥下界的渐近方差协方差矩阵。

A.7 随机数量阶

使用关于变量序列收敛速度的有用记号是，利用记号(O,o)或者大 O、小 o 记号表示序列的数量阶。

如果 $\lim(a_N/g(N))$是有限非零的，非随机实数的序列 a_N 称为$O(g(N))$的；而如果 $\lim(a_N/g(N))$是 0，a_N 称为$o(g(N))$的。因而，a_N 是$O(g(N))$的，如果它具有与函数 $g(N)$相同的数量阶；而 a_N 是$o(g(N))$的，如果它具有比 $g(N)$较小的数量阶。例如，$(3/N)+(5/N^2)$是 $O(1/N)$或者 $O(N^{-1})$，因为对于大 N 来说，其特性像一个常值时间 N^{-1}，并且是 $o(N^{-1/2})$的，却比 $o(N^{-1})$大。

这种记号可被推广到随机变量序列的随机数量阶上。这类记号变为(O_p,o_p)。

定义 A.22(随机数量阶)：随机变量序列 b_N 称为 $O_p(g(N))$的，如果：

$$0<\operatorname{plim}\frac{b_N}{g(N)}<\infty$$

而随机变量序列 b_N 称为 $o_p(g(N))$ 的，如果：

$$\operatorname{plim}\frac{b_N}{g(N)}=0$$

绝大多数时候，对某一个常值 c 来说，$g(N)=N^{-c}$。估计量 $\hat{\theta}$ 关于 θ_0 是一致的，这能够被写成 $\hat{\theta}=\theta_0+o_p(1)$，因为它等于 θ_0 加上一个依概率趋于 0 的项。估计量$\hat{\theta}$ 关于 θ_0 是根号 N 一致的，能够被写成$\hat{\theta}=\theta_0+O_p(N^{-1/2})$，从而 $N^{-1/2}(\hat{\theta}-\theta_0)=O_p(1)$。

A.8 其他一些结果

本节包括有限样本的条件期望以及期望与变换进行交换的一些重要结果。

定理(期望迭代定律)：对于随机变量 Y 与 X：

[1] 原著中该式为 $\mathcal{N}[\mu,\sigma^2]$，应该为 $\mathcal{N}[0,\sigma^2]$，这里已改。——译者注

$$E[Y]=E_X[E_{Y|X}[Y|X]]$$

其中，E[·]表示 Y 的无条件或边际均值的期望，$E_X[\cdot]$表示关于 X 的边际 cdf 的无条件期望，而 $E_{Y|X}[\cdot|X]$表示给定 X 时关于 Y 的条件分布的条件期望。

这个结果意味着，如果我们首先获得给定 X 时 Y 的条件均值，然后针对 X 取期望值，那么将获得 Y 的无条件均值。参见拉奥（Rao，1973，第 97 页）的证明。例如，若 $E[u|\mathbf{x}]=0$，那么 $E[u]=E_x[E[u|\mathbf{x}]]=E_{\mathbf{x}}[0]=0$。

定理（方差分解）：对于随机变量 Y 与 X：

$$V[Y]=E_X[V_{Y|X}[Y|X]]+V_X[E_{Y|X}[Y|X]]$$

其中，$V[Y]$表示 Y 的无条件方差，$E_X[\cdot]$表示关于 X 的边际 cdf 无条件期望，$V_{Y|X}[Y|X]$表示给定 X 时 Y 的条件方差，$V_X[\cdot]$表示关于 X 的无条件分布的方差，$E_{Y|X}[\cdot|X]$表示给定 X 时 Y 的条件分布的条件期望。

总之，Y 的无条件方差等于：(1)（针对 X）条件方差的期望值与(2)（针对 X）条件均值的方差之和。记住该关系式的简单方法是，清楚认识到，无条件方差等于 EV 加上 VE。参见拉奥（Rao，1973，第 97 页）的证明。

定理（詹森不等式）：如果 Z 是一个随机变量，使得 $E[Z]$存在，并且 $g(\cdot)$是一个凸函数，那么：

$$g(E[Z])\leqslant E[g(Z)]$$

然而，如果 $g(\cdot)$ 是一个凹函数，那么：

$$g(E[Z])\geqslant E[g(Z)]$$

对于非线性模型来说，这个结果极为重要，它已由拉奥（Rao，1973，第 58 页）证明。该定理强调了平均个体的特性与平均特性之间的差异。例如，假定一个指数模型合适，满足 $E[y|\mathbf{x}]=\exp(\mathbf{x}'\boldsymbol{\beta})$。于是，由于该指数函数是凹的，故詹森不等式蕴含着 $\exp(E[\mathbf{x}'\boldsymbol{\beta}])\geqslant E[\exp(\mathbf{x}'\boldsymbol{\beta})]$。因此，在个体平均特性处计算的条件均值 $\mathbf{x}=E[\mathbf{x}]$大于无条件均值 $E[y]=E[E[y|\mathbf{x}]]=E[\exp(\mathbf{x}'\boldsymbol{\beta})]$。

A.9 文献注释

一个带有证明的经典文献来源于拉奥（Rao，1973，第 108～130 页），这里，我们尽可能地引用其结论。所概括的结果还密切依赖于雨宫（Amemiya，1985，第 3 章）以及怀特（White，2001a）的书。

研究生水平的教科书，诸如格林（Greene，2003）的书提供了对重要结果的总结。更为高等的教科书包括戴维森和麦金农（Davidson and MacKinnon，1993）、亨德里（Hendry，1995）、鲁德（Ruud，2000）以及伍德里奇（Wooldridge，2002）的书，这些书提供了至少与本书同样详细的处理。戴维森（Davidson，1994）为经济计量学家提供了随机理论的一个深入详细的研究。尤其在使用斯卢茨基定理及克拉默定理时，前面提到的术语会因参考文献不同而表现得不一样。

B 伪随机采样

在这个附录里,我们阐述重要的一元分布的密度或概率质量函数以及前二阶矩,然后表述从这些分布中生成随机采样的一些方法。

表 B.1 连续随机变量的密度与矩[a]

随机变量	pdf $f(x)$	均值;方差
一致 $\mathcal{U}[a,b]$	$1/(b-a)$	$\frac{(a+b)}{2}$; $\frac{(a-b)^2}{12}$
正态 $\mathcal{N}[\mu,\sigma^2]$	$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ; σ^2
指数 $\mathcal{E}[\lambda]$	$\lambda e^{-\lambda x}$, $\lambda>0$	$1/\lambda$; $1/\lambda^2$
伽玛 $\mathcal{G}[a,b]$	$\frac{1}{\Gamma(a)b^a}x^{a-1}e^{-\frac{x}{b}}$	ab; ab^2
贝塔 $\mathcal{B}[a,b]$	$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$	$\frac{a}{a+b}$; $\frac{ab}{(a+b)^2(a+b+1)}$
逻辑斯蒂 $\mathcal{L}[a,b]$	$e^{-\frac{x-a}{b}}\Big/\left[b(1+e^{-\frac{x-a}{b}})^2\right]$, $-\infty<a<\infty$	a; $(b\pi)^2/3$
卡方 $\chi^2(n)$	$\frac{x^{n/2-1}e^{-x/2}}{\Gamma(n/2)2^{n/2}}$	n; $2n$
t $t(v)$	$f(x)=\frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)\sqrt{v\pi}}\left(1+\frac{x^2}{v}\right)^{-\frac{v+1}{2}}$	0; 当 $v>2$ 时, $\frac{v}{v-2}$
F $F(w,v)$	$f(x)=\frac{\Gamma\left(\frac{w+v}{2}\right)(v/w)^{v/2}}{\Gamma\left(\frac{w}{2}\right)\Gamma\left(\frac{v}{2}\right)}x^{w/2-1}\times\left(x+\frac{v}{w}\right)^{-\frac{w+v}{2}}$	当 $v>2$ 时, $\frac{v}{v-2}$ 当 $v>4$ 时, $\frac{2v^2(v+w-2)}{w(v-4)(v-2)^2}$

[a] 所有参数限制如下:对于一致分布变量,$b>a$;对于正态分布变量,μ 无限制,$\sigma^2>0$;对于指数分布变量,$\lambda>0$;对于伽玛分布变量,$a,b>0$;对于贝塔分布变量,$a,b>0$;对于逻辑斯蒂分布变量,a 无限制且 $b>0$;对于 t 分布变量,v 为整数;对于 F 分布变量,v 与 w 都必须为整数。

表 B.2　连续随机变量生成器

随机变量	变量范围	随机变量生成器
一致 $\mathcal{U}[a,b]$	$a \leqslant x \leqslant b$	$x=a+(b-a)r,\ r\sim\mathcal{U}[0,1]$
正态 $\mathcal{N}[\mu,\sigma^2]$	$-\infty < x_1, x_2 < \infty$	$\begin{cases} x_1=\mu+\sigma\sqrt{-2\ln(r_1)}\cos(2\pi r_2) \\ x_2=\mu+\sigma\sqrt{-2\ln(r_1)}\sin(2\pi r_2) \end{cases}$
$[r_1, r_2\sim\mathcal{U}[0,1]$;所得到的数对于 x_1 与 x_2 是独立随机变量。]		
指数 $\mathcal{E}[\lambda]$	$0 \leqslant x < \infty$	$x=-\frac{1}{\lambda}\ln(r)$
伽玛 $\mathcal{G}[a,b]$	$0 \leqslant x < \infty$	$\begin{cases} \text{(i)}\ x=-\frac{1}{\lambda}\ln(\prod_{i=1}^{a} r_i) \text{ 或 } x=\sum_{i=1}^{a} E_i \\ \text{(ii)}\ x=-\frac{1}{\lambda}[\ln(\prod_{i=1}^{m} r_i)-y_1 y_2] \end{cases}$
(i) $r_i\sim\mathcal{U}[0,1]$; a 是整数,E_i 是 iid 指数随机变量 当 $a=1$ 时,得到指数随机变量 (ii) a 是非整数,$a=m+q$, $0<q<1$, $m=$整数 y_1, y_2 是独立的 $\mathcal{B}(q,1-q)$与 $\mathcal{E}(1)$		
贝塔 $\mathcal{B}[a,b]$	$0 \leqslant x \leqslant 1$	$\begin{cases} \text{(i)}\ x=y_1/(y_1+y_2) \\ \text{(ii)}\ x=r_1^{\frac{1}{a}}\Big/\left(r_1^{\frac{1}{a}}+r_2^{\frac{1}{b}}\right),\ \left(r_1^{\frac{1}{a}}+r_2^{\frac{1}{b}}\right)\leqslant 1 \end{cases}$
(i) a,b 都是整数,y_1 服从 $\mathcal{G}(k,a)$, y_2 服从 $\mathcal{G}(k,b)$ k 可任意选取 (ii) a,b 都是非整数, $r_i\sim\mathcal{U}[0,1]$;连续生成数对 r_1 与 r_2, 直到 $\left(r_1^{\frac{1}{a}}+r_2^{\frac{1}{b}}\right)\leqslant 1$		
逻辑斯蒂 $\mathcal{L}[a,b]$	$-\infty < x < \infty$	$x=a+b\ln\left(\frac{r}{1-r}\right)$
$[r\sim\mathcal{U}[0,1]]$		
卡方 $\chi^2(n)$	$0 \leqslant x$	$\sum_{i=1}^{n} y_i^2$
[n 是一个整数;y_i 是独立的 $\mathcal{N}(0,1)$]		
t $t(v)$	$-\infty < x < \infty$	$x=y_1/\sqrt{y_2/v}$
[y_1 服从 $\mathcal{N}(0,1)$;y_2 与 y_1 独立,y_2 服从 $\chi^2(v)$]		
F $F(w,v)$	$0 \leqslant x$	$x=(y_1/w)/(y_2/v)$
[y_2 与 y_1 独立,分别服从 $\chi^2(v)$与 $\chi^2(w)$]		

表 B.3　离散随机变量概率质量函数与矩

随机变量[a]	pmf $f(x)$	均值；方差
二项 $Bi[n,p]$	$\binom{n}{x}p^x(1-p)^{n-x}$	np；$np(1-p)$
泊松 $\mathcal{P}[\lambda]$	$e^{-\lambda}\lambda^x/x!$	λ；λ
负二项 $NB[n,p]$	$\binom{n+x-1}{x}p^n(1-p)^x$	$\frac{n(1-p)}{p}$；$\frac{n(1-p)}{p^2}$

[a] 对于二项随机变量来说，$0\leqslant p\leqslant 1$ 且 n 为正整数；对于泊松随机变量来说，$\lambda>0$；对于负二项随机变量来说，$0<p<1$，$n>1$。

表 B.4　离散随机变量生成器

随机变量	取值范围	随机变量
二项 $Bi(n,p)$	$x=0,1,\cdots,n$	设 $x=0$； 做 n 次循环 　生成一致分布$[0,1]$上的 r 　当 $r\leqslant p$ 时，则令 $x=x+1$ 输出 x
泊松 $\mathcal{P}(\lambda)$	$x=0,1,\cdots$	设 $x=0$；$t=0$ 做 n 次循环，直到 $t<\lambda$ 　生成指数随机变量 y 　令 $t=t+y$ 　$x=x+1$ 输出 x
负二项 $NB(n,p)$	$x=0,1,\cdots$	由 $\mathcal{G}\left(n,\frac{1-p}{p}\right)$生成$\lambda$ 由 $\mathcal{P}(\lambda)$ 生成 x 输出 x

［Ｇｅｎｅｒａｌ Ｉｎｆｏｒｍａｔｉｏｎ］
书名＝微观经济计量学　方法与应用
作者＝（美）卡梅伦，（美）特里维迪著
页数＝８５４
ＳＳ号＝１２６５０４３８
出版日期＝２０１０．０７
出版社＝上海财经大学出版社

书名
前言
目录
序言